esc 300$00

ISSN 0

técnica

evista de Engenharia

ublicação da Associação dos Estuda

Instituto Superior Técnico

INSTITUTO TECNOLÓGICO
PARA A EUROPA COMUNITÁRIA

APOIAMOS A INOVAÇÃO TECNOLÓGICA

CRIAMOS "DESENVOLVIMENTO"

- FORMAÇÃO AVANÇADA EM ÁREAS ESTRATÉGICAS
- INVESTIGAÇÃO & DESENVOLVIMENTO
- ORGANIZAÇÃO E PLANEAMENTO DE PROJECTOS COMUNITÁRIOS
- PRESTAÇÃO DE SERVIÇOS PARA A INDUSTRIA
- ESTÁGIOS INTERNACIONAIS PARA RECÉM-LICENCIADOS
- INCUBAÇÃO DE EMPRESAS

Av. Rovisco Pais • 1096 LISBOA CODEX • Telef. 803660 - ext. 1538/1657 • Fax 899242

Técnica - revista de Engenharia

Propriedade: Associação dos Estudantes do Instituto Superior Técnico
Número de Pessoa Colectiva: 501055606

Director Científico: Manuel de Abreu Faro
Director Executivo: Paulo David Simões
Coordenação Administrativa: Gerardo Vieira Lisboa

Direcção, Administração e Publicidade:
Associação dos Estudantes do Instituto Superior Técnico
Av. Rovisco Pais 1000 Lisboa Portugal
Telefones directos: 8481018 - 8489323
Extensão IST: 1248 - 1260
Fax: 899242 (a/c AEIST)
E-Mail (bitnet, EARN): aeistdoc@eta.ist.rccn.pt

Composição e impressão: Editorial Império, Limitada
R. do Salitre 155 - 2 1296 Lisboa Codex

Registo DGCS 101407 Depósito Legal 41683 ISSN 0040-1714

Tiragem: 5000 exemplares / Publicação Trimestral
Preço de capa: 300$00 Assinatura anual (Portugal): 1000$00
(Estrangeiro): USD$12

Contribuiram para este número:
Junta Nacional de Investigação Científica e Tecnológica
Conselho Directivo do Instituto Superior Técnico

Técnica - número único de 1990 Impressão: Fevereiro de 1991

EDITORIAL 3

Artigo Convidado

Situação Científica sobre a Compreensão da Língua Natural
Helder Coelho 5

Notas Técnico-Científicas

Combustão de Resíduos Florestais - Alguns Resultados
J. M. P. Ventura, E. C. Fernandes e D. F. G. Durão 21

Solução de um Problema Indeterminado: a Inversão de Dados Sísmicos
João Duarte Fonseca 27

The ISFET in Medicine
C. A. C. Sequeira 39

The First Principle of Thermodynamics and the Non-Separability of the Quantities "Work" and "Heat"
Rodrigo de Abreu 43

Dual Galilei Transforms
M. J. Martins 47

Os artigos publicados são da inteira responsabilidade dos respectivos autores.

EDITORIAL

Num passado não distante a TÉCNICA configurou-se numa revista plena de interesse, integrando trabalhos de vanguarda nos domínios da ciência e da técnica.

Espera-se que a partir de 1991 retome o ritmo perdido.

Caminhar para esses objectivos e dar realidade às grandes finalidades de uma revista com a tradição da TÉCNICA é tarefa que apela necessariamente para os docentes, investigadores e alunos do I.S.T. e bem assim para todos aqueles que cultivam os domínios científicos e técnicos que servem a Engenharia.

Para relançar a revista há que acreditar na TÉCNICA, na acção que pode desenvolver na comunidade técnico-científica e, nomeadamente, ter presente que, em princípio, os artigos da TÉCNICA são referidos nos *Abstracts* das diferentes especialidades científicas. Assim tem acontecido.

Repetimos quanto, de essencial, se julgou dever salientar no número anterior. No entanto, alguma coisa mais há que dizer:

Assim, é-nos grato confirmar que o ritmo que se imprimiu à TÉCNICA se mantém, não esmoreceu.

É-nos profundamente grato, também, salientar a forma generosa como o Conselho Directivo do I.S.T. tem acarinhado os nossos esforços, ouvindo-nos e concedendo-nos significante apoio material.

Situação Científica sobre a Compreensão da Língua Natural

Helder Coelho

Professor Catedrático do Instituto Superior de Economia e Gestão,
Universidade Técnica de Lisboa
Chefe de projecto do INESC

RESUMO

O avanço muito rápido e a popularização recente da Inteligência Artificial (IA) obrigam-nos a realizar com regularidade pontos de situação sobre os avanços e os recuos verificados, não só nas suas áreas de aplicação, mas também no que respeita às suas metodologias, técnicas e instrumentos. A compreensão da língua natural (LN) é um bom exemplo para analisarmos as fronteiras actuais da IA, pois esta disciplina tem progredido por saltos e num terreno multidisciplinar.

No presente artigo focam-se os seus aspectos centrais, as dificuldades presentes e sobretudo as esperanças em conhecer melhor os modos de processar os significados das frases trocadas entre os interlocutores em presença, quer isoladas ou em contexto. Em síntese, trata-se de falar sobre o tema geral da comunicação homem-máquina, mas sob o ponto de vista dos fenómenos cognitivos subjacentes.

ABSTRACT

The fast advance and recent popularization of Artificial Intelligence (AI) force us to do regular states of the art about the application areas, methodologies, techniques and tools. The understanding of natural language is a good example for the analysis of the actual borders of AI, because the discipline progressed by jumps and in a multidisciplinary ground.

Along the present paper we focus the central issues, the present difficulties, and the hopes in knowing better the processing modes of sentence meanings exchanged among interlocutors, either isolated or in context. Briefly, we discuss the general topics of human-machine communication, taking the point of view of the cognitive associated phenomena.

1 - INTRODUÇÃO

Após um período dourado na década de 70 (perguntas/respostas SHRDLU deWinograd, tradução automática de Wilks, perguntas/respostas LUNAR de Woods, compreenção de histórias MARGIE, SAM e PAM de Schank, pesquisa de bases de dados ROBOT de Harris e LIFER de Hendrix), a área da compreensão da língua natural (LN) da Inteligência Artificial (IA) entrou num período longo de amadurecimento, vislumbrando-se apenas algumas saídas cujo êxito possa ser calculado com facilidade.

Ao longo deste artigo enunciam-se as fronteiras que se colocam contra o progresso acelerado da I&DE desta área, apontando-se as várias vias de investigação em curso e esboçando os terrenos mais férteis para o avanço e as dificuldades que o impedem.

Mas, o que é compreender uma língua natural? Como se realiza uma tal compressão? E, onde e quando se faz essa compreensão? Estas perguntas continuam a motivar respostas que ainda são parciais e que necessitam de ser articuladas, como se resume na figura 1. Por um lado, tais interrogações têm permitido a descoberta de problemas cujas soluções implicam o recurso a diversas disciplinas. E, por outro lado as dificuldades surgidas têm aberto lugares de intervenção cientifica e têm alargado as fronteiras do nosso conhecimento.

A LN é um termo impreciso, dependente da clarificação das capacidades linguísticas e das capacidades humanas (memória, raciocínio, resolução de problemas, formação de hipóteses, classificação, planeamento, consciência social e aprendizagem). A LN, como afirma William Woods, assume a compreensão da parte do ouvinte, e não se restringe a uma mera descodificação. Compreender uma LN envolve não só o conhecimento da estrutura da língua (léxico, sintaxe, semântica e morfologia), o conhecimento de uma situação (contexto), mas também o conhecimento geral (mundo e universo) e o recurso ao raciocínio. Uma tal compreensão pode ser automatizada ao longo de diversos

---

Este texto corresponde à lição convidada, proferida pelo autor durante o 1. Congresso Ibero-Americano de Inteligência Artificial (IBERAMIA-88) que decorreu de 11 a 13 de Janeiro de 1988 em Barcelona (Espanha).

Original entregue para publicação em 2 de Maio de 1988
Revisto a 16 de Fevereiro de 1991

Área:	Compreensão da LN	
Objectivo:	Aprender a construir programas de computador que possam comunicar e interagir com os seres inteligentes num modo em tudo idêntico ao adoptado pelos próprios seres humanos no seu inter-relacionamento.	
Motivação:	Alargar a nossa compreensão sobre a natureza da linguagem e do pensamento.	
Sub-áreas: relacionadas	Compreensão da LN (escrita ou falada) Tradução automatizada da LN Geração de textos Lexicografia e morfologia	Frases isoladas Diálogos, Discursos

figura 1 - Quadro resumo

níveis como se indica na figura 2, mas ainda não está claro em que nível podemos afirmar que finalmente atingimos a compreensão da LN.

DESCRIÇÃO GRAMATICAL ("PARSING")
Processamento de frases em descrições estruturais usando descrições gramaticais da estrutura linguística.

INTERPRETAÇÃO SEMÂNTICA
Aplicação da descrição estrutural da frase numa forma lógica que apresenta o significado independente do contexto.

INTERPRETAÇÃO CONTEXTUAL
Aplicação das formas sintáctica e lógica na representação final da compreensão da frase.

REPRESENTAÇÃO DO SIGNIFICADO
Escrita numa linguagem que suporta uma componente de inferência para modelar o raciocínio natural.

figura 2 - Níveis de processamento da compreensão da LN

A leitura da figura 2 permite desde logo evocar **dois sub-problemas** que estão interligados e que são de facto os obstáculos maiores para se atingir uma compreensão plena:

Representação: Como são representadas as possíveis interpretações diferentes em cada nível?

Interpretação: Como são produzidas as representações apropriadas em cada nível, desde a entrada até esse nível?

A compreensão da LN é muito difícil e cada vez envolve mais esforço para que se verifique um progresso real no seu conhecimento. Nos últimos anos os avanços têm sido lentos e de espectro cada vez mais lato, porque esta área envolve muito mais do que responder a perguntas isoladas, tal como se aponta na figura 3, ao apresentar-se o campo de intervenção do Center for the Study of Language and Information (CSLI) criado em 1983 nos EUA, através da Stanford University, do SRI International e da Xerox Parc.

A título de exemplo podemos ver como três destas disciplinas citadas na figura 3 se referem à língua e como a encaram com pontos de vista diferentes, mas complementares:

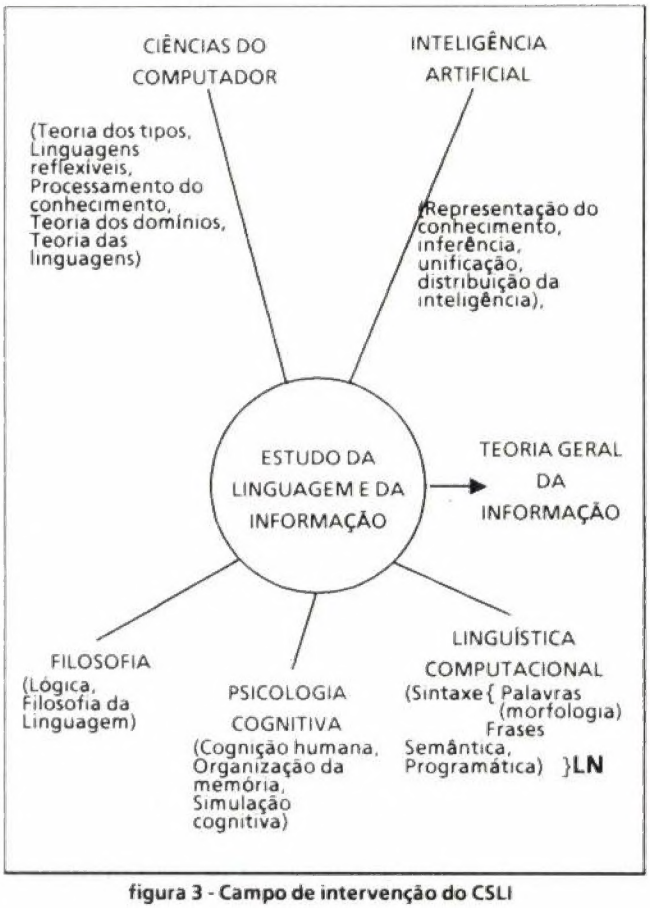


figura 3 - Campo de intervenção do CSLI

LINGUÍSTICA:
modelos estruturais da língua

PSICOLOGIA COGNITIVA:
modelos do uso da língua

INTELIGÊNCIA ARTIFICIAL:
modelos (computacionalmente eficientes) do uso da língua em comunicação.

Seleccionando a disciplina Linguística Computacional podemos enumerar os seus vários objectivos, através da sua decomposição nas suas três áreas tradicionais:

ANÁLISE SINTÁTICA	Gramática de contexto livre (CFG) Gramática de contexto livre aumentado Constrições gramaticais Fontes de ambiguidade sintática
ANÁLISE SEMÂNTICA	Representação do significado Constrições semânticas Análise de quantificadores
ANÁLISE DO DISCURSO	Identificação de informação implícita Estabecimento da coerência textual Enquadramentos Guiões

e observar como elas se cruzam entre si. Em resumo, podemos afirmar que estamos perante fenómenos de uma grande complexidade, e cuja compreensão exige uma acção multidisciplinar em várias frentes.

A nível internacional e após a publicação da tese de doutoramento de Terry Winograd em 1972 o progresso real na LN foi significativo, com contribuições importantes, principalmente nos domínios da sintaxe e da semântica. A área passou de 2º lugar (1981) para 1º lugar (1983) no que respeita o número de comunicações apresentadas, na Conferência Internacional Conjunta de IA (IJCAI), que se realiza de dois em dois anos. Em 1985 desceu para 3º lugar, confirmando-se esta queda em 1987 ao passar para 5º lugar [6].

Do ponto de vista mais geral, a situação da **comunicação homem-máquina** tem sido até hoje menos que satisfatória. Esta área é muito importante para o alargamento da população dos utilizadores de computadores, pois tem como objectivo oferecer modos mais flexíveis de interacção, facultando aos não informáticos interfaces amigáveis com os produtos informáticos. No entanto, e face à grande expectativa, 1) não existe ainda uma norma para as representações internas, 2) nem todos os fenómenos linguísticos podem ser resolvidos ao mesmo nível de processamento, 3) existem dificuldades em coordenar estratégias de análise que são múltiplas e independentes, e 4) é difícil descobrir-se uma solução geral para a descrição do significado das frases por causa da natureza da estrutura gerada. Tudo isto reflecte-se nas capacidades reais das interfaces, em geral eficientes apenas em mundos fechados e perante pequenos subconjuntos de uma LN.

Observando também o que é um **agente inteligente** descobrimos imediatamente cinco traços que são pertinentes para os trabalhos em compreensão da LN. A saber, a comunicação (leia-se a facilidade de conversa), o conhecimento interno (do sistema), o conhecimento do mundo (e a faculdade de aquisição de mais conhecimentos), a intencionalidade (a elaboração e a compreensão de planos/objectivo), e a criatividade (uma nova maneira de olhar para algo que muda o nosso mundo). Alguns destes aspectos são focados nas secções seguintes, mas as dificuldades do progresso têm sido enormes levando os investigadores a ensaiarem soluções alternativas e provisórias na procura de uma solução geral e global. A diversidade de metodologias, técnicas, formalismos e teorias demonstra que estamos ainda longe de conseguir uma visão completa dos fenómenos linguísticos.

A compreensão da língua natural pode ser discutida essencialmente através de dois pontos de vista, o reconhecimento ao nível da frase e os diálogos interactivos, que serão em seguida discutidos com algum detalhe.

2 - Reconhecimento ao nível da frase

Quando se aborda o reconhecimento ao nível da frase é necessário ter presente que uma língua é um veículo para transmitir conceitos de mudança, de lugar, de causalidade e de objectivo. E, que a compreensão de uma língua natural recorre a diversos dispositivos, tais como a referência pronominal (anáfora), a elipse, a modificação da cláusula relativa, a quantificação, a modificação de conceitos por adjectivos e advérbios, e as transformações focadas na atenção. Assim, existem **características** que um utilizador dos sistemas de LN não deve esquecer. Mas, ao procurar identificá-las é necessário saber colocar as perguntas apropriadas:

Pergunta: errada
- Será que o sistema X tem a característica Y?

Perguntas: certas
- Até que ponto da característica Y é capaz o sistema X?
- Até que ponto é importante a característica Y para a minha aplicação?

No entanto, nem todas as características estão actualmente compreendidas e dominadas, como em seguida se anota.

C1: **COBERTURA**:
competência linguística de um sistema, analisada ao longo das dimensões seguintes:

Cobertura Lexical:
quantidade de vocábulos (preposições, conjunções, artigos, etc.)

Cobertura sintáctica:
fenómenos sintácticos (formas verbais complexas, cláusulas relativas, formas de perguntas, passivos, comparativos, cláusulas subordenadas, advérbios de tempo e lugar, expressões de medida, elipse, pronomes e conjunções)

Cobertura semântica:
capacidade de compreensão

C2: **HABITALIDADE**:
velocidade de reconhecimento e adaptação do utilizador às limitações do sistema

C3: **INFERÊNCIA**:
capacidade de tirar conclusões com base nos dados e no conhecimento disponível

C4: **CONTROLE DA INTERPRETAÇÃO PELOS UTILIZADORES**:
capacidade de recorrer ao contexto (com limitações)

C5: **USO DE PRONOMES**:
referência a acções e a objectos no discurso (com limitações)

C6: **OUTROS TIPOS DE REFERÊNCIA**:
recurso a agregados

C7: **ELIPSE**:
capacidade de encher partes da conversa através do contexto

C8: **QUANTIFICAÇÃO**:
uso de conhecimento de senso comum para interpretar palavras como alguns e quaisquer (com limitações)

C9: **NEGAÇÃO**:
em frases nominais e verbais, e combinada com a quantificação (com limitações)

C10: **TEMPO E ASPECTO**:
em aberto e investigação, pois não existem mecanismos gerais para representar efectivamente e eficientemente acontecimentos e objectivos que se alteram com o tempo (em aberto)

C11: **CONJUNÇÃO E DISJUNÇÃO**:
recurso aos "e" e "ou"; em combinação com as elipses e outros fenómenos (em aberto)

C12: **ENTRADA TELEGRÁFICA**:
recurso a abreviaturas (em aberto)

C13: **ENTRADAS NÃO GRAMATICAIS**:
recurso a formas habituais (em aberto)

O problema básico do processamento da LN é a resolução da ambiguidade (sintáctica ou estrutural, semântica das palavras, casos, referencial e literalidade), isto é pretende-se traduzir frases ambíguas em representações internas não ambíguas. Contudo, não existe ainda uma norma geral para as representações internas e são variados os tipos de representações de conhecimento disponíveis na área da IA que têm sido até agora experimentados:

TIPOS DISPONÍVEIS DE REPRESENTAÇÕES

- EXPRESSÕES DE UMA LINGUAGEM DE PESQUISA (para acesso a BD's)
- FORMULAS LÓGICAS
- ÁRVORES DE DESCRIÇÃO COM PALAVRAS EM NÓS TERMINAIS (para tradução automatizada)
- EXPRESSÕES LISP (para sistemas periciais)
- INSTANCIAÇÕES DE ENQUADRAMENTOS (vários)
- DEPENDÊNCIAS CONCEPTUAIS (para compreensão de histórias)

Estes tipos de representações adequam-se a certos fenómenos linguísticos ou a aspectos particulares da compreensão de uma LN, e portanto revelam-se incompletos quando se pretende construir um sistema global.

Quanto aos formalismos gramaticais existem actualmente cerca de trinta propostas, e nenhuma delas é capaz de dar uma resposta global, completa e categórica aos fenómenos linguísticos. Vejamos algumas variedades desses formalismos, agrupados em torno da sintaxe e da semântica.

TEORIAS SINTÁTICAS:

GBT: GOVERNMENT BINDING THEORY (Chomsky, 1981)

DCG: DEFINITE CLAUSE GRAMMAR (Coelho, 1979; Pereira e Warren, 1981)

FG: FUNCTIONAL GRAMMAR (Kay, 1979)

FUG: FUNCTIONAL UNIFICATION GRAMMAR (Kay, 1984)

LFG: LEXICAL FUNCTIONAL GRAMMAR (Bresnan, 1982)

GPSG: GENERALIZED PHRASE STRUCTURE GRAMMAR (Gazdar, 1985)

LG: LEXICON GRAMMAR (Gross, 1986)

CG: CATEGORIAL GRAMMAR (Dowty, 1982; Bach, 1983; Steedman, 1985)

IG: INDEXED GRAMMAR (Gazdar, 1985)

HG: HEAD GRAMMAR (Pollard, 1984)

HPSG: HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR (Sag e Pollard, 1987)

IG: INCREMENTAL GRAMMAR (Kempen, 1987)

CFQG: CONTEXT-FREE QUEUE GRAMMAR (Manaster-Rena, 1986)

TAG: TREE ADJOINING GRAMMAR (Joshi, Levy & Takahashi, 1975)

TEORIAS SEMÂNTICAS:

SITUATIONAL SEMANTICS (Bairwise e Perry, 1983)

DISCOURSE REPRESENTATION THEORY (Kamp, 1981; Heim, 1982)

PROCEDURAL SEMANTICS (Wood, 1968)

CONCEPTUAL DEPENDENCY (Schank, 1975)

MONTAGUE SEMANTICS (Montague, 1970)

COMPOSITIONAL SEMANTICS

KNOWLEDGE-BASED SEMANTICS

Alguns destes formalismos/teorias possuem aspectos semelhantes e portanto podem convergir, do modo como é indicado em seguida na figura 4.

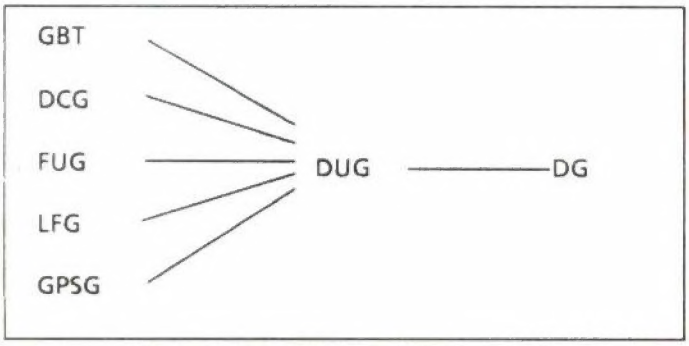


Figura 4 - Convergência de formalismos

As razões para uma tal convergência destes formalismos gramaticais são as seguintes:

1) Apoiam-se em relações de dependência, recorrendo a noções como "cabeça",
2) Tomam em consideração os aspectos funcionais,
3) Defendem que a sintaxe deve ser restringida do ponto de vista lexical, e por isso colocam uma grande parte da informação gramatical no léxico, e
4) Baseiam os seus algoritmos no princípio da unificação, que diz que "As categorias complexas são colocadas em acordo no contexto sintáctico".

Vejamos em detalhe as características de três destes formalismos convergentes:

LFG

Baseia as suas c-estruturas sintácticas no princípio da estrutura sintagmática.

FUG

As f-estruturas funcionais representam relações entre functores e argumentos.

GPSG

Introduz mais regras para restringir de forma adequada o formalismo.
O controle da análise é distribuido por muitas fontes.

Os **formalismos gramaticais** podem-se também agrupar em **famílias canónicas**, tais como:

- "DEFINITE CLAUSE GRAMMAR" (DCG)
- "FUNCTIONAL UNIFICATION GRAMMAR" (FUG)
- "LEXICAL FUNCIONAL GRAMMAR" (LFG)
- "GENERALIZED PHRASE STRUCTURE GRAMMAR" (GPSG)
- "UNIFICATION-BASED GRAMMAR" (PATR, PATR-II, S-PATR)

as quais comunicam entre si como se mostra na figura 5

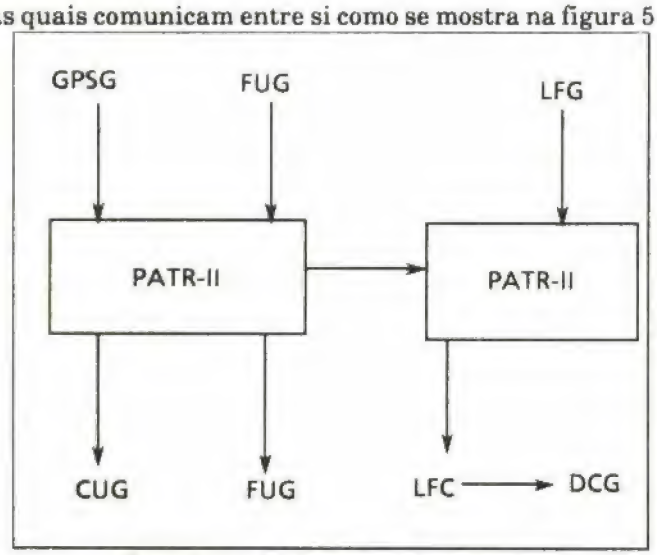


Figura 5 - Famílias canónicas de gramáticas

Consideremos dois exemplos que mostram como estes formalismos comunicam entre si:

Exemplo 1:

Análise:	"Dowty gapping analysis"
Método:	CUG
Gramática de implementação:	PATR-II

Exemplo 2:

Análise:	GP & S análise de verbos auxiliares
Método:	LFG
Gramática de implementação:	DCG
Parsing:	via tradução BUP ("Bottom Up Parser")

Vejamos agora na figura 6 as **técnicas** que foram sendo propostas para a análise da LN, isto é para a tradução das fases da LN em apenas uma representação interna.

- ASSOCIAÇÃO DE PADRÕES	(ELIZA/PARRY)
- DESCRIÇÃO CONDUZIDA SINTÁCTICAMENTE	(ATN's)
- GRAMÁTICAS SEMÂNTICAS	(SOPHIE, LIFER)
- INSTANCIAÇÃO DE ENQUADRAMENTOS DE CASOS	(ELI)
- ESPERE E VEJA	(MARCUS)
- PERITO EM PALAVRAS	(SMALL)
- CONECCIONISMO	(SMALL)
- RASPAGEM	(FRUMP, IPP)

Figura 6 - Técnicas de análise de LN

Destas técnicas de análise vejamos algumas delas com mais pormenor:

ASSOCIAÇÃO DE PADRÕES (Weizenbaum, Colby)

- interpretação global das frases, através da associação de padrões (interpretações) das palavras à frase de entrada.

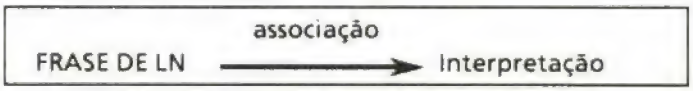


Figura 7 - Esquema da associação de padrões

DESCRIÇÃO CONDUZIDA SINTACTICAMENTE (Bobrow, Fraser, Woods)

- interpretação construtiva através da combinação da estrutura (unidades de alto nível: frases,cláusulas, sintagmas) e do significado das palavras ou de outros constituintes de nível mais baixo.

  A sua ineficiência pode ser combatida através da mistura da descrição ("parsing") com a interpretação.

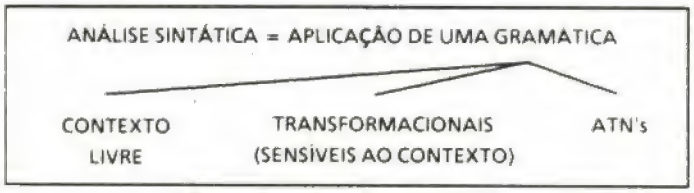


Figura 8 - variedades sintácticas

GRAMÁTICAS SEMÂNTICAS (Burton, Hendrix)

- interpretação construtiva, através da combinação de categorias definidas do ponto de vista sintáctico e semântico (reconhecimento por uma gramática sintáctica antes da interpretação semântica).

  Desvantagens: tendem a tornar-se muito grandes e são difíceis de construir

  Variantes:
  1)- misturar sintaxe com semântica (RUS/Bobrow)
  2)- abstracção dos aspectos comuns linguísticos da gramática semântica para uma certa aplicação (TEAM/Grosz)
  3)- multi-estratégias de descrição para gerar formas canónicas antes do uso da gramática semântica (DYPAR/Carbonell)

INSTANCIAÇÃO DE ENQUADRAMENTOS DE CASOS (Simmons, Schank, Riesbeck)

- natureza recursiva e capacidade para combinar o reconhecimento de constitutintes chaves via da base para o topo (BUP) com a instanciação de constituintes menos estruturados do topo para a base (TDP).
  Vantagens: descrição robusta adequada a enfrentar erros de escrita do utilizador.

Ora, nem todos os fenómenos linguísticos podem ser resolvidos no mesmo nível de análise, e daí a existência de diferentes técnicas.

Exemplo:

Correcção de erros	: nível léxical
Elipse	: níveis sintáctico e pragmático
Anáfora	: níveis semântico e sintáctico

Perante estas dificuldades Carbonell e Hayes têm defendido que se devem adequar as estratégias de análise aos tipos de construção linguística específicas. Esta via está ainda em investigação, apresentando desde já uma grande dificuldade: coordenação de estratégias de análise que são múltiplas e independentes.

As fases e a ordem na análise das frases de LN também podem ser variadas consoante se pretende enfrentar certos fenómenos linguísticos. Em seguida apresentam-se três alternativas, e citam-se os autores que se adoptaram.

A) Sintaxe e depois Semântica ELISÃO
ANÁFORA
QUANTIFICAÇÃO

(Woods, 1972)
(Mellish, 1979)
(Pereira, 1983)
(Grosz et al, 1987)

B) Semântica e depois Sintaxe ANÁFORA

(Schank & Tesler, 1969)
(Wilks, 1972)
(Riesbeck & Schank, 1976)

C) Sintaxe mais Semântica QUANTIFICAÇÃO
MODIFICADORES

(Woods, 1970)
(Burton, 1970)
(Winograd, 1972)
(Dahl, 1978)
(Coelho, 1979)

Também não existe uma direcção e estilo único de processamento da compreensão dos significados, como se faz notar em seguida na figura 9.

	NÃO DETERMINISTICOS
ANALISADORES	
	DETERMINISTICOS
	DO TOPO PARA A BASE (TDP)
ANALISADORES	OLHAR PARA A FRENTE (LOOK HEAD)
	DA BASE PARA O TOPO (BUP)

Figura 9 - Direcções de processamento

Tudo isto aponta para grandes dificuldades em encontrar uma solução geral quanto ao reconhecimento de frases de LN, por causa da natureza de estrutura gerada. Na figura 10 apresenta-se a organização de uma interface transportável (Grosz et al, 1987), onde se pode observar a complexidade das diversas fases de processamento.

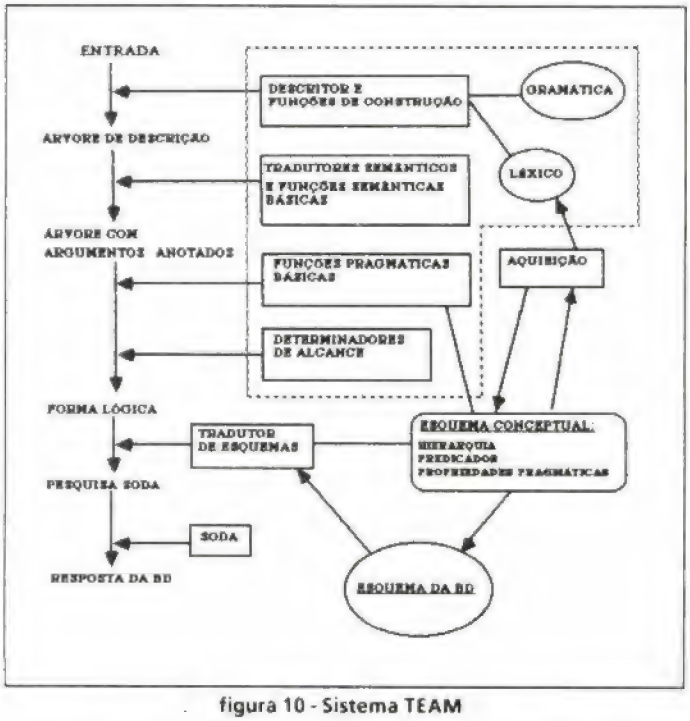


figura 10 - Sistema TEAM

A exploração de instrumentos, técnicas e metodologias tem sido contínua ao longo dos anos, como

se aponta na figura 11, de forma esquemática e para as variedades lógicas.

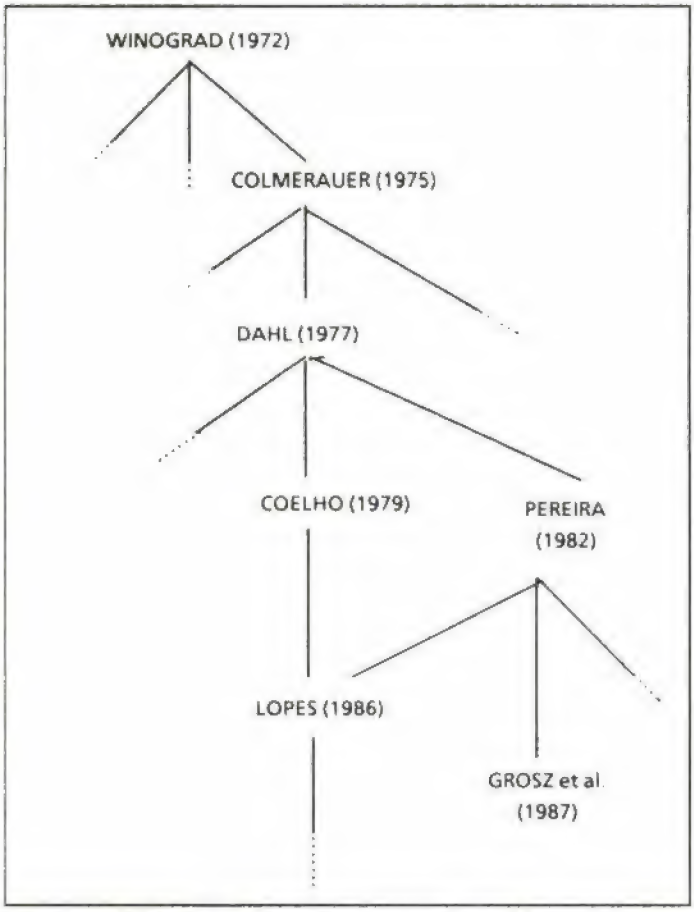


figura 11 - Evolução da investigação em LN

Consideremos agora o tópico das formas lógicas para compararmos algumas destas vias de investigação, nomeadamente a de (Coelho, 1979) com o sistemaTUGA, a de (Warren e Pereira, 1981) com o sistema CHAT-80, a de (Lopes, 1986) com o sistema SSIPA e a de (Grosz et al, 1987) com o sistema TEAM. Para cada um destes sistemas apresentamos em seguida uma frase de LN e a correspondente descrição lógica do seu significado.

TUGA > Nenhum editor tem outro autor.

```
não(
    para(X1,
        e(editor (X1)
            para(X2,
                e(autor (X2),
                    tem (X1,X2)),
                card (X2, maior, 0))),
            card (X1, maior, 0)))
```

Chat-80 > Which countries bordering the Mediterranean border Asian countries?

```
answer (C) <=    country (C) &
                 border (C, mediterranean) &
                 exists (C1, country (C1) &
                 asian (C1) & borders (C,C1))
```

Na figura 12 comparam-se dois sistemas, o CHAT-80 e o SSIPA, e mostra-se como a evolução do primeiro para o segundo foi obtida através da maior complexificação do

SSIPA > Escreva uma mensagem !

```
imperativa <= =
    eu (A:X) &
    mensagem (B:Y, de C:Z, para : D:W) &
    escrever (A.X, mensagem & B:Y)
```

Chat-80 > Escreva uma mensagem !

```
imperativa <= =
    eu (X) & mensagem (Y,Z,W) & escreva (X,Y)
```

processamento, logo gerando uma maior capacidade de compreensão.

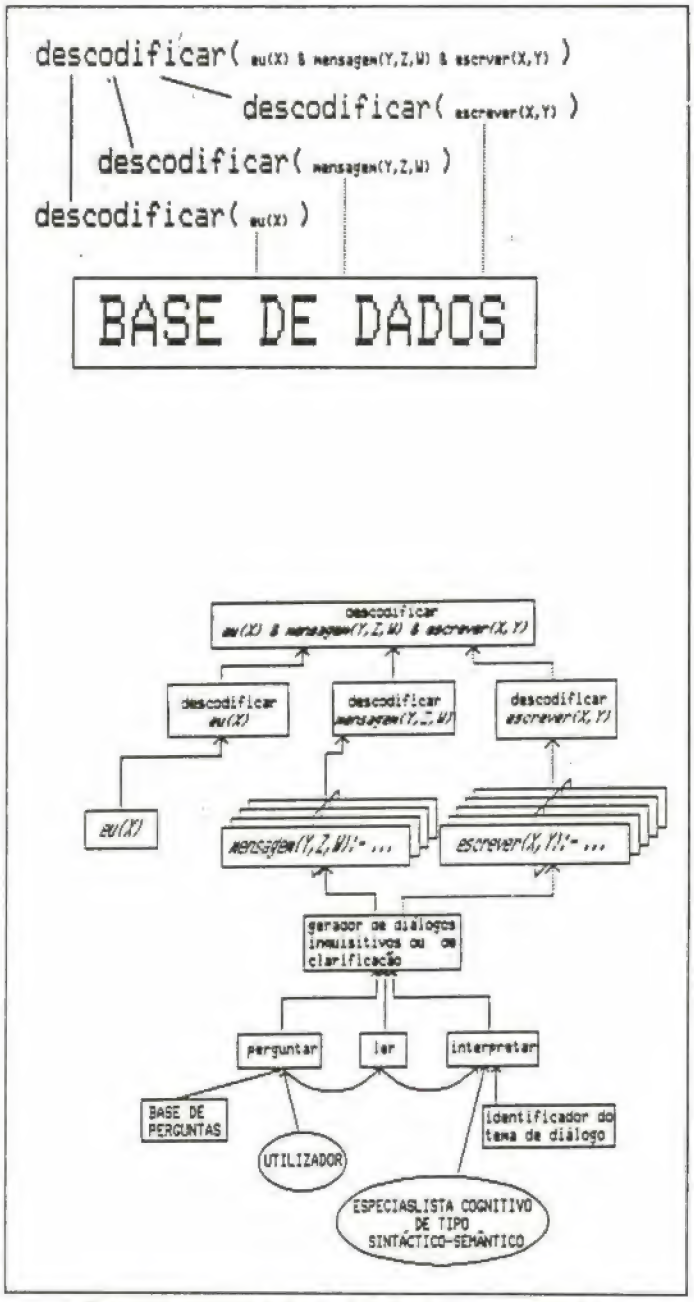


figura 12 - Comparação de dois tipos de compreensão

No caso do sistema TEAM o tipo de representação é ligeiramente diferente, como se mostra em seguida.

Voltemos de novo à questão da representação dos significados e vejamos através de uma série de exemplos

```
TEAM > Show each continent's highest peak

SHOW EACH CONTINENT GEN EST HIGH PEAK.

FOUR PARSES WERE FOUND.

(IMPERATIVE
    (ALL + WORLDC-CONTINENT + 2
        (+ WORLD-CONTINENT + +
         + + WORLDC-CONTINENT + 2)
        (THE PEAK + 1
            (AND
                (+ PEAK + + PEAK + 1)
                (((*SUPER*(*MORE**HIGH))
                    + PEAK + 5
                    (AND (+ PEAK + + PEAK + 5)
                        (*PK CONT-CONTINENT-OF
                        PEAK + 5 + WORLDC-CONTINENT + 2)))
                    + PEAK + 1))
            (*SHOW + YOU + + SPEAKER + + PEAK + 1))))
    FOR EVERY CONTINENT
        FOR THE UNIQUE PEAK
        SUCH THAT THE PEAK IS THE HIGHEST PEAK SUCH THAT
            THE CONTINENT IS CONTINENT OF THE PEAK
        YOU SHOW SPEAKER THE PEAK
```

como (Allen, 1987) organiza as descrições ao longo de dois níveis, o da sintaxe e o da semântica.

Sintaxe: estrutura da frase

John sold the book to Mary.

```
(S      SUBJ      (NP NAME   John
                      NUM    {3s})

MAIN-V    sold
TENSE     {PAST}
VOICE     {ACTIVE}
OBJ       (NP DET  the
              HEAD  book
              NUM   {3s})
MODS      (PP PREP  to
              POBJ  (NP NAME  Mary
                        NUM   {3s}))
```

Semântica: forma lógica

Jack took Mary to the Halloweenball.

```
(PAST il TAKE-ACTION  [AGENT (NAME p1 PERSON "Jack")]
                      [THEME(NAME p2 PERSON "Mary")]
                      [TO-LOC(DEF/SING b1 PARTY
                          (CELEBRATES b1 HALLOWEEN))])
```

Se focarmos a forma de representação de (Allen, 1983) podemos observar como a anáfora entre frases é resolvida.

When does the flight from Boston arrive in Chicago?

```
(S   WH-QUERY   (PP1 HEAD    when)
     SUBJ (NP1  DET  the
                HEAD   flight
                NUM    {3p}
                MODS   (PP  PREP    from
                            POBJ    (NP2 NAME Boston)))
     MAIN-V     arrive
     TENSE      PRES
     MODS       ((PP PREP    in
                     POBJ    (NP3 NAME Chicago))
                  →PP1))
```

Os referentes são organizados numa estrutura que contém a história da interacção, e que serve depois para apoiar a resolução de problemas linguísticos.

Constituintes	Objecto	Tipo
NP3	CHI	CITY
NP1	FLIGHTO124	FLIGHT
NP3	BOS	CITY

Consideremos agora a pergunta seguinte para explicar como é resolvida a referência introduzida por "that flight".

Does that flight serve lunch?

```
••• that flight •••

(NP DET    that
    HEAD   flight
    NUM    {3s}
    SEM    (PRO/SING f1 FLIGHT))
```

O teste sobre a lista com a história da interacção é feito recorrendo à informação da abertura SEM daquela estrutura. Em primeiro lugar, ocorre um falhanço com o constituinte 1, e depois dá-se um sucesso com o constituinte 2.

REF = FLIGHT 0124

/

referente para "that flight"

Vejamos, finalmente, como se constrói a representação final de uma frase segundo (Allen, 1987). Consideremos a seguinte frase,

Jack bought a car.

```
(PAST b1 BUY [AGENT (NAME j1 PERSON "Jack")]
             [THEME(INDEF/SING C1 CAR)])
```

1) **Aplicação** dos sintagmas nominais aos termos da base de conhecimentos (BC) que representam os seus referentes

referente de Jack : JACK 1
referente de a car : CAR 18

2) **Aplicação** das partes proporcionais (acções) da forma lógica à BC

definição da classe PURCHASE

```
(B1 isa PURCHASE with
    [Buyer JACK1]
    [Seller Seller (B1)]
    [Object CAR 1]
    Precondition: HAS(JACK 1, Prince (CAR 1))
    Delete: HAS(JACK 1, Prince (CAR))
    Add: HAS(JACK 1, CAR 1)
    Decomposition: GOTO(JACK 1, Loc(Seller (B1)))
                   GIVE(JACK 1, Seller(B1),
                                (CAR 1))
                   GIVE(Seller(B1), Jack 1,
                                CAR 1))
```

Cria uma instância P1 coloca o papel Buyer em JACK 1 e o perfil objecto em CAR 18

```
(P1  isa PURCHASE with
     [Buyer JACK 1]
     [Seller Seller (P1)]
     [Object CAR 18])
```

Ao longo destes exemplos evidenciou-se a questão central da representação do conhecimento do significado de uma frase de entrada de um sistema pergunta-resposta, tendo em vista a selecção de uma resposta adequada, a qual pode ser realizada em dois passos sucessivos:

1) resposta à pergunta como procura na base de dados, e
2) geração de uma resposta que ajudará melhor o objectivo do utilizador, nomeadamente para:
   - antecipar as suas próximas acções, por exemplo realizar um acto que ocorreria posteriormente embora não tivesse ainda sido explicitado, e
   - inspecionar os planos reconhecidos em busca de obstáculos, isto é as partes do plano que o utilizador não seria capaz de realizar por falta de conhecimento ou por causa de crenças erradas.

3. Diálogos interactivos

Nas conversas ordinárias de todos os dias, os participantes fazem hipóteses sobre o conhecimento que vão trocando, assinalam as suas hipóteses através das pressuposições pragmáticas das suas frases, e negoceiam incompreensões sobre o conhecimento mútuo que vão desenvolvendo. Tudo isto é feito de modo rotineiro com o bom senso e através de maquinarias que não conhecemos ainda com rigor. Sabemos apenas que as conversas se constroem perante um conjunto de fenómenos do discurso, os quais podem ser simples e complexos, como em seguida se enumeram na figura 13.

Ao ensaiarmos resolver a compreensão de um discurso/conversa temos de entrar em linha de conta não só com aqueles fenómenos, mas também com uma grande variedade de aspectos, dos quais se destacam os seguintes:

- ELIPSES	SIMPLES
- ANÁFORAS	
- SINTAGMAS NOMINAIS DEFINIDOS	
- FRASES EXTRAGRAMATICAIS	
- FRASES METALINGUÍSTICAS	COMPLEXOS
- ACTOS DE DISCURSO INDIRECTO	

figura 13 - Fenómenos do discurso

- unidades de significado dependentes e não isoladas
- contexto: • colocação das frases no mundo real
  • ligação linguística entre frases
- regras da estrutura do discurso: como as frases podem ser combinadas para constituir diálogos
- papel do conhecimento de senso comum para derivar relações semânticas entre constituintes do diálogo
- procura de intenções e propósitos das frases

Daqui, pode-se concluir que existem quatro aspectos fundamentais que devemos abordar de modo mais profundo quando pretendemos automatizar os diálogos interactivos: a estrutura do discurso, a semântica do discurso, os actos de discurso e pragmática, e as modalidades de entrada/saída.

3.1 Estrutura do discurso

A compreensão da estrutura do discurso/conversa envolve os seguintes objectivos:

- Compreender as **relações** entre as suas partes,
  - frases
  - segmentos/cursos
  - monólogos
  - explicações
  - diálogos
  - tácticas: trocas
  - estratégias/planos

- A **estrutura do discurso** afecta a interpretação das frases individuais: define o contexto semântico que pode ser invocado para determinar a interpretação de um pronome, uma descrição definida ou um termo descritivo.

- **Métodos para determinar a estrutura** do discurso que estão associados à compreensão de estruturas linguísticas e retóricas:
  - mecanismos sintácticos de narrativa
  - semântica e sintaxe das frases
  - sequências pergunta/resposta
  - operadores de discurso (OK, sim, não, está bem)
  - convenções de divertimento
  - encaixe de discursos e convenções de retorno

Recentemente, as linhas de investigação desenvolvem-se em duas frentes:

capturar o significado, ou as relações de coerência entre cláusulas e segmentos, as quais funcionam em conjunto para comunicar ideias

interligadas (Hobbs, 1979 e 1985; Mann e Thompson, 1983)

compreender as relações estruturais em discursos não coerentes, os quais são dominados por interrupções, hesitações, etc.
(Reichman, 1981 - Teoria do espaço do contexto, Grosz e Sindner, 1986 - Teoria das estruturas do discurso, Polanyi e Scha, 1984 - Modelo do discurso dinâmico)

3.2 Semântica do discurso

Quanto à semântica do discurso, existe uma tese que orienta actualmente os trabalhos de investigação.

TESE: as condições de verdade de uma frase podem ser descritas matemáticamente como uma função de estados em valores da verdade

Existem várias operações relacionadas com o processamento de uma frase num discurso, a saber:

- determinar o significado da frase no contexto,
- integrar o significado da frase com o significado do discurso, e
- actualizar o contexto que será usado para interpretar a próxima frase.

A interpretação de uma frase pode ser alcançada através da inferência dos objectivos conversacionais, e dos papeis sociais dos participantes. No entanto, é necessário recorrer a regras de expectativa sobre o discurso, tais como aquelas que:

- operam nos níveis sintáctico, semântico e pragmático para resolver problemas linguísticos, como a elipse, e as que
- permitem focar a atenção do sistema sobre o diálogo.

3.3 Actos de discurso e pragmática

Quanto aos actos de discurso e pragmática defrontam-se neste momento duas linhas de investigação exemplificadas na figura 14.

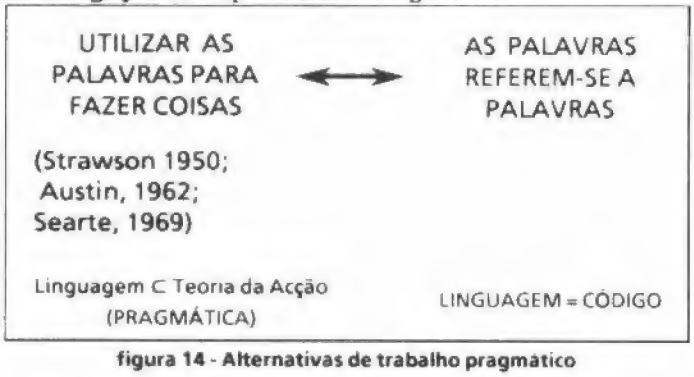


figura 14 - Alternativas de trabalho pragmático

As teses em presença são as seguintes:

1) O **significado** de uma mensagem linguística é apenas representado pelo conteúdo.
2) A **atribuição de intenções** ao falante deve ser uma parte integral do processo de compreensão do ouvinte.
3) Uma **teoria de compreensão** da linguagem deve determinar até que ponto as estratégias que as pessoas usam para chegar a explicações satisfatórias do comportamento físico dos outros podem ser empregues na compreensão dos actos de discurso.

Compreender o discurso requer inferir intenções e crenças que conduzem o locutor a produzir um comportamento. E, em geral, a comunicação em LN envolve muito mais do que responder a perguntas isoladas. O recurso ao contexto é de facto fundamental, como se indica na figura 15.

Figura 15 - Modelos

Eis algumas das linhas de investigação empreendidas desde 1980:

- compreender actos indirectos (Perraut, 1980)
- fornecer respostas que ajudem (Allen,1980)
- interpretar perguntas mal formadas (Carberry,1986)
- compreender elipses entre frases (Allen, 1980; Carberry, 1985)
- identificar o tipo de respostas pretendido (Sidner, 1983)

A modelação do conteúdo de uma conversa/história é feita basicamente através de duas vias, indicadas na figura 16,

Figura 16 - Modelação da conversa

as quais podem ser resumidas através de três aspectos:

Características:	
nada de não habitual ocorre	coisas não esperadas ocorrem
Representação:	
baseada em acções	baseada em acções
Processamento:	
associação de padrões	composição do plano com sub-planos

Em resumo, em oposição estão:

GRANDES DESCRIÇÕES	contra	PEQUENAS ACÇÕES + SISTEMA GERAL DE RACIOCÍNIO para agrupar as acções em estruturas de planos

Uma língua não pode ser completamente compreendida sem o recurso ao conhecimento de todos os dias que os falantes têm acerca do mundo. Para conversar um sistema deve compreender as ligações entre pares de frases e tirar conclusões adequadas. E, daí a necessidade de recorrer às crenças dos agentes em presença. Logo, as representações do conhecimento devem ter capacidades para incluir crenças. A distinção dos diferentes significados pretendidos é feita recorrendo aos modelos de crenças, onde se insere a informação sobre o que o falante/agente acredita. Existem várias representações possíveis:

1) Etiquetar cada facto da BC com o nome do agente que acredita nesse facto.
2) Dividir a BC em bases separadas, designadas por espaços de crença, cada uma representando um conjunto de crenças de um agente particular.
3) Cada espaço de crenças inclui diferentes acções e estruturas de planos de forma a contemplar os planos individuais dos agentes, com:
   - crenças encastradas
   - crenças mútuas/partilhadas
   - crenças sobre os conhecimentos e habilidades dos outros

A comunicação é modelada de diversos modos, entre os quais se realçam os trabalhos pelos seguintes autores:

(Grice,1957; Austin,1962; Searle,1969)
(Cohen & Perrault,1979; Allen & Perrault,1980)
(Grosz & Sidner,1986; Litman & Allen,1987)

A identificação do significado pretendido numa conversa é o objectivo a alcançar, como se indica na figura 17.

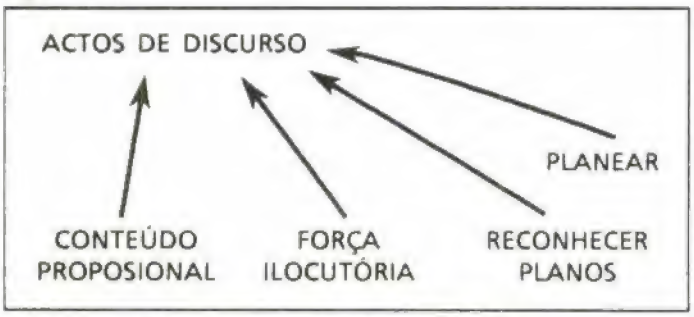


figura 17 - Identificação do significado

A definição das circunstâncias apropriadas sob as quais são realizados os actos do discurso dependem das condições sobre as crenças e intenções do falante e do conhecimento sobre as convenções da linguagem.

Os modelos de diálogo do utilizador que actualmente estão disponíveis são:

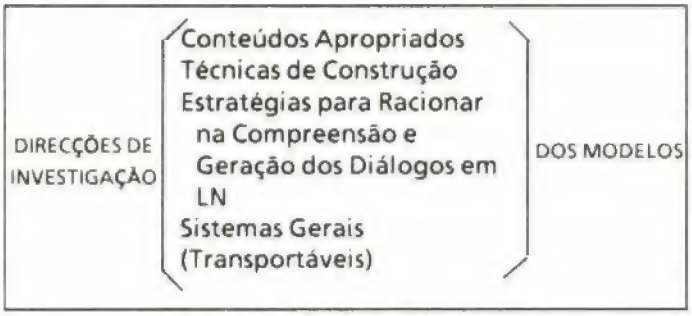


figura 18 - Modelos de diálogo

Vejamos, a título de exemplo, alguns dos projectos em curso na RFA que abordam a temática dos actos de discurso:

- HAM-ANS : HAMBURG APPLICATION - oriented NATURAL LANGUAGE SYSTEM
  - sistema de reserva de hoteis
  - sistema de análise de sequências de imagens
  - sistema de BD relacionais
- XTRA: EXPERT TRANSLATOR
  - preenchimento de impressos de impostos
- VITRA : VISUAL TRANSLATOR
  - interface para um sistema de visão
- SC : SINIX CONSULTANT
  - sistema de ajuda em LN para a SINIX
- WISBER : Wisenbarierter Beratungsdialog
  - sistema para consulta financeira

3.4 Modalidades de entrada/saída

As modalidades de entrada/saída de um sistema de compreensão da LN podem desdobrar-se do modo seguinte:

1) Oral vs. escrita
2) Grau de interacção
3) Frequência espacial
4) Frequência temporal
5) Possibilidade de comunicação para a linguística
6) Referentes concretos
7) Especificidade da audiência
8) Separação dos participantes

Obviamente, cada sistema sobredetermina apenas uma ou algumas destas modalidades, diminuindo assim as dificuldades de automatização dos aspectos subjacentes.

4 - Problemas da LN em investigação

A análise das Actas das últimas Conferências Internacionais sobre Linguistica Computacional (COLING 86, Bona) e Inteligência Artificial (IJCAI 87, Milão) permite agrupar as áreas onde se trabalha mais intensamente.

No campo da Linguística Computacional, as áreas são: a semântica, os diálogos, a gramática, o software, a tradução automatizada, a descrição/análise, o conhecimento, a morfologia, a geração e o discurso. No campo da Inteligência Artificial estas áreas estão também presentes e os problemas em investigação podem-se destacar de uma lista exaustiva:

- Problemas principais

Desambiguação do sentido das palavras descrição/análise, verificação da consistência do discurso, resolução de termos equívocos, resolução da anáfora, violação de constrições categoriais entre predicados e argumentos, construção de acordos (referências), e geração incremental de frases.

- Problemas secundários

Construção dinâmica do modelo de uma tarefa/plano, reconhecimento de planos em situações, erro de comunicação em referência, classificação dos diálogos por negociação, coordenação da compreensão de diversos tipos de discurso incluindo pedaços de informação cujo nível de granularidade é diferente, modelação dinâmica do uso da língua, ajuste das descrições dos objectos do conhecimento do utilizador, reparação do modelo do utilizador face à disparidade de planos, selecção de guiões, e manutenção do estado corrente do guião revelante.

A título de exemplo, seleccionemos a parte da compreensão de LN do projecto japonês da quinta geração, e vejamos através da sua situação em 1987 como alguns daqueles problemas estão presentes.

As direcções de trabalho apontavam para os seguintes objectivos:

1) Sistema para a compreensão do discurso (DUALS)
2) Caixa de ferramentas para a linguagem (LTB)
3) Gramática de estruturas sintagmáticas generalizadas japonesas (JPSG)

O caminho de investigação japonês procura desde o início encontrar respostas para a seguinte questão central:

Pergunta: O que é **compreender** uma LN?

através da realização de um sistema concreto, denominado DUALSI, e com as seguintes características:

1984 : DUALSI

- análise sintáctica com BUP ("Bottom-up parser")
- análise de contexto com identificação de objectos baseada em Prolog (resolução das elipses e anáforas)
- análise de discursos
- resolução de problemas com fins específicos
- escrito em LFG ("Lexical Functional Grammar") e CIL ("Complex Indeterminate Language"/"Situation Semantics Theory")
- baseado no paradigma da programação lógica

Os principais passos da evolução do projecto japonês, no que respeita a compreensão da LN, são os seguintes:

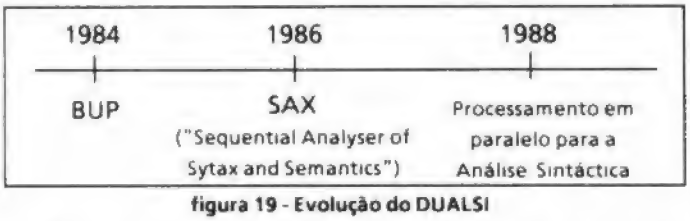


figura 19 - Evolução do DUALSI

Também aqui neste projecto de investigação se nota uma tendência de fazer convergir três formalismos gramaticais num único formalismo, o CIL:

e o argumento principal para uma tal fusão baseia-se na articulação de vários formalismos gramaticais baseados na unificação.

Do ponto de vista das teorias da LN adoptadas destacam-se as seguintes:

- ACTOS DO DISCURSO
- PLANOS-OBJECTIVO
- CÁLCULOS DE EVENTOS
- ESTRUTURAS DE DISCURSO/SITUAÇÕES

que estão sendo trabalhadas, de forma a serem posteriormente integradas.

A questão fundamental da investigação japonesa consiste em descobrir como organizar estas teorias de forma a que elas cooperem entre si. O objectivo imediato traduz-se em como organizar a estrutura do discurso com a Teoria das Situações, isto é como mudar a estrutura complexa do discurso em uma estrutura do topo para base. E, o objectivo futuro desdobra-se em várias tarefas: dicionários de grandes dimensões (centenas de milhares de palavras), compreensão do falante (quem fala), compreensão das referências e condução de raciocínio de senso comum. A esperança dos investigadores japoneses baseia-se no amadurecimento da Teoria dos Discursos. Deve-se ressaltar que algumas destas tarefas estão também a serem resolvidas num outro programa de investigação ambicioso, o da MCI nos EUA.

É interessante observar que no projecto japonês se podem observar algumas das grandes questões da investigação actual a nível internacional. Eis uma delas:

Pergunta: Em que nível podemos afirmar que finalmente atingimos a compreensão da LN?

Figura 20 - Em direcção à compreensão

Os japoneses trabalham assim no espaço entre a descrição e a compreensão total, e tentam avançar algumas ideias numa tentativa de obter respostas para aquela pergunta.

5 - Aplicações da LN

O campo de aplicação da compreensão da LN é multifacetado, sendo a **geração de interfaces** bastante popular para a construção e engenharia de diversos sistemas:

- sistemas de gestão de bases de dados
- sistema de apoio à decisão
- sistemas periciais e outros baseados em conhecimento
- pacotes gráficos
- sistemas tutores inteligentes
- sistemas pergunta/resposta
- sistemas de ajuda inteligentes

No caso das interfaces dos Sistemas de Gestão de Bases de Dados (SGBD's) há uma visão simplista do contexto. Existe uma base de dados fixa com conhecimentos que não mudam, e a interpretação de qualquer base é virtualmente das frases precedentes. Quer isto dizer, que não existe contexto conversacional de facto.

A elaboração de sistemas pergunta/resposta, quer para os SGBD's ou para os sistemas tutores, é um domínio privilegiado para se estudarem os aspectos básicos em compreensão e geração de LN. De facto, ao abordarem-se os tipos de perguntas, estamos a estudar o apoio à interacção através da geração de outros tipos de respostas, tais como respostas correctas, respostas úteis, respostas que não enganem, e respostas indirectas.

O estudo do apoio à interacção passa também pela compreensão de diversas funções, tais como a definição de termos, as paráfrases, a correcção de equívocos, e a modificação de perguntas.

As interfaces espertas permitem aumentar a tolerância aos erros de dactilografia (ou de pronúncia) e às divergências gramaticais. Donde, a necessidade de capacidade de reparação de erros!

Vejamos a definição de termos, onde se pretende atingir a adaptação à terminologia do utilizador, quer através da aquisição de informação ou da geração de explicações.

Existem quatro métodos:

1) O utilizador equaciona o novo termo ou reconstrói-o com termos já conhecidos pelo sistema;

2) O utilizador é guiado passo a passo através de um processo de definição capaz de fornecer ao sistema informação sintáctica e semântica acerca do novo termo;

3) O sistema assume que o termo desconhecido é um valor da base de dados, e espera pelo resultado da procura;

4) O utilizador é solicitado a escrever de novo o seu pedido com outros termos.

E a investigação actual espraia-se em três direcções:

- adaptar as definições ao contexto do discurso corrente,
- adaptar as definições ao conhecimento e/ou aos objectivos prévios do utilizador,e
- recorrer tanto a imagens como a textos.

No que respeita a aquisição do conhecimento, destacam-se neste momento três direcções de trabalho, a saber:

Verbos	(Grosz et al, 1987)
Campos de aspectos	(Grosz et al, 1987)
Regras de gramática	(Viccari, 1987)

ao longo de dois projectos, um nos EUA e outro em Portugal:

projecto TEAM	(Grosz, Appelt, Martin, Pereira, 1987)
projecto LNEC	(Viccari, 1987)

A motivação destes objectos é a obtensão de maior flexibilidade das interfaces, e o domínio dos processos de:

- geração de perguntas, e de
- extracção de informação sintáctica e semântica das respostas.

Numa **conversa** a maior parte das dificuldades provêm do facto dos participantes terem visões

diferentes acerca do mundo, e ou não descrevem este facto, ou falham alguma coisa contra isso. Daí, o aparecimento de equívocos (disparidades entre as crenças dos dois participantes) e a necessidade de os corrigir.

O objectivo desta correcção consiste na identificação e na reparação dos equívocos.

Exemplos de sistemas já construidos:

CO-OP	(Kaplan, 1982)
	(Mercer, 1984)
	(Gal & Minker, 1985)

O estado da arte da construção de interfaces pode resumir-se ao longo de dois aspectos dominantes:

para fins específicos:

- Dificuldade ou mesmo impossibilidade de serem feitas alterações significativas ao sistema por alguém que não seja projectista.
- Duração de construção: de dois meses a cinco anos para os programadores experientes.

para fins gerais:

- Tempo de desenvolvimento muito longo (alguns anos) para produzir componentes independentes do domínio.
- Esforço de aplicação a um domínio particular: semanas ou meses.

Vejamos alguns exemplos de produtos desenvolvidos ao longo dos últimos dez anos. Em interfaces de LN para bases de dados, LUNAR, REQUEST, TQA, REL, PLANES, NL MENU SYSTEM, EXPLORER e RUS. E, no que respeita os produtos comerciais, destacam-se INTELLECT, THEMES, RAMIS, CLOUT, PLUME e EXPLORER.

Os tutores inteligentes é uma área de aplicação da IA onde se torna necessário a aprendizagem da língua natural, para que eles se comportem inteligentemente quando dialogam com os seus alunos, quer através da análise das suas frases ou da geração de respostas ou de perguntas adequadas. Ora, aprender uma língua é uma tarefa necessária para estudar como se interactua numa língua, pois isto implica simultaneamente resolver três problemas, nomeadamente a análise gramatical, as interpretações semânticas e pragmáticas, e o raciocínio sobre os objectivos dos alunos. Estes problemas que se encontram ligados entre si, jazem no seio dos temas processamento e língua natural e projectam-se contra o espaço referencial da representação do conhecimento, pois da selecção de um formalismo depende em grande parte a procura das melhores soluções para construir sistemas capazes de conversar com os seres humanos.

Finalmente, a participação activa dos sistemas em diálogo com os seus utilizadores depende fundamentalmente não só de maiores conhecimentos linguísticos de ordem sintáctica-semântica, mas também de ordem pragmática, os quais se situam fora do contexto do trabalho em IA. Com eles, e com a participação e colaboração de linguístas, filósofos, psicólogos e informáticos o progresso será mais forte. Em Portugal, não tem sido fácil até agora congregar estes esforços. O exemplo do ILTEC (Instituto de Linguística Teórica e Computacional) é encorajador.

6 - Conclusão

Os aspectos que se encontram correntemente **em investigação em LN** são: semântica formal, reconstrução lógica, descrição gramatical como dedução, negação e disjunção, complexidade gerativa forte, unificação de termos vs. grafos, sistemas categóricos combinatórios, partilha de estrutura, restrição e dispositivos não monotónicos. Mas, são os aspectos da descrição gramatical em paralelo e das formas de organização do léxico que se constituem como merecedores de grande atenção, pois da sua cabal resolução poderão surgir progressos nesta área.

O estado da arte da compreensão da LN em 1987 pode ser sintetizada ao longo dos seguintes **vectores:**

- necessidade de grandes quantidades de conhecimento do mundo real
  - novos formalismos de representação
  - novas técnicas de inferência para desambiguar as frases
  - novas organizações de memória
  - novos modos de codificação

  A IA não atingiu ainda o estádio em que consegue manipular de forma rotineira o conhecimento necessário! Daí, a variedade de sistemas piloto ou de protótipos com um subconjunto de LN e com poucos exemplos de demonstração...

- necessidade de capacidade de aquisição de conhecimento
  - o sistema deve ter uma grande quantidade de conhecimento antes de começar a aprender

- necessidade de aprendizagem da LN pelos sistemas que conversam
- necessidade de capacidade de detecção e resolução dos erros e das incompreensões do utilizador
- necessidade de analisar cada domínio de aplicação para pré-codificar inferências específicas.

Em Portugal, os trabalhos de investigação sobre compreensão da LN iniciaram-se em 1976 no Centro de Informática do LNEC, em redor de Luís Moniz Pereira e Helder Coelho. Actualmente, estes trabalhos prosseguem através de projectos autónomos no INESC, na UNL e na FCUP.

Bibliografia Sumária

[1] Allen, J.F. (1987). *Man-machine dialogues*, Tutorial 7, I JCAI.

[2] Amarel, S. (1987). *Problem solving, in Encyclopedia of Artificial Intelligence*, S. C. Shapiro (ed.), John Wiley & Sons.

[3] Carbonell, J.G., Hayes, P.J. (1987). *Natural-language interfaces, in Encyclopedia of Artificial Intelligence*, S.C. Shapiro (ed.), John Wiley & Sons.

[4] Coelho, H. (1980). *Elementos da Engenharia da linguagemTese para Especialista*, LNEC.

[5] Coelho, H. (1980). *A program conversing in Portuguese providing a library service, Universidade de Edimburgo*, 1979; 2ª edição , LNEC.

[6] Coelho, H. (1990). *Inteligência Artificial: balanço da década de 80*, Editorial Caminho.

[7] Filgueiras, M.C.O. (1984). *Compreensão da linguagem natural por computador: uma metodologia*, Universidade Nova de Lisboa.

[8] Grosz, B.J., Applet, D.E., Martin, P.A. Pereira, F.C.N. (1987). *TEAM: an experiment in the design of transportable natural-languages interfaces*, Artificial Inteligence 32.

[9] Henschen, L. (1987). *Reasoning, in Encyclopedia of Artificial Intelligence*, S.C. Shapiro (ed.), John Wiley & Sons.

[10] Jacob, P.S. (1987). *Knowledge-intersive natural language generation*, Artificial Intelligence, 33.

[11] Lopes, J.G.P. (1986). *Conceptualização de um sistema interlocutor automático*, Universidade Técnica de Lisboa.

[12] Pereira, F.C.N. (1983). *Logic for natural language analysis*, SRI International, Technical Note 275.

[13] Scha, R.J., Bruce, B.C., Polanyi, L. (1987). *Discourse understanding, in Encyclopedia of Artificial Intelligence*, S.C. Shapiro (ed.), John Wiley & Sons.

[14] Waldrop, M.M. (1984). *Natural language understanding, Science*, **224**, 27 April.

[15] Weber, B. (1987). *Question answering, in Encyclopedia of Artificial Intelligence*, S.C. Shapiro (ed.), John Wiley & Sons.

Apêndice 1:

INTERVENÇÃO PORTUGUESA NA ÁREA DE LN

Teses de doutoramento

Coelho, H. (1979)
A program conversing in Portuguese providing a library service.
Universidade de Edimburgo (realizada no LNEC e orientada pelo doutores David Warren e Luís Moniz Pereira); 2ª edição, LNEC, 1980.

Filgueiras, M.C.O. (1984)
Compreensão da linguagem natural por computador: uma metodologia.
Universidade Nova de Lisboa (realizada na UNL e orientada pelo doutor Luís Moniz Pereira).

Lopes, J.G.P. (1986)
Conceptualização de um sistema interlocutor automático
Universidade Técnica de Lisboa (realizada no LNEC e orientada pelo doutor Helder Coelho).

Experiências práticas (programas aplicados)

CALCC - *Consulta automática da legislação da construção civil*
(J. C: Cotta, LNEC, 1980)

Combustão de Resíduos Florestais - Alguns Resultados

J. M. P. Ventura, E. C. Fernandes e D. F. G. Durão

Departamento de Engenharia Mecânica, Instituto Superior Técnico

RESUMO

Para uma aplicação correcta dos modelos existentes de propagação de fogos às condições portuguesas, é necessário obter informação sobre as características de queima dos combustíveis florestais mais correntes em Portugal, que permita fazer as necessárias adaptações ao sub-modelo do combustível constante desses modelos. É este o objectivo de um projecto em curso no Departamento de Engenharia Mecânica do IST.

O presente trabalho descreve a instalação e instrumentação utilizada e mostra alguns resultados obtidos nesta área. Foram feitos ensaios de queima de leitos de agulhas de pinheiro, com vento transversal e sem vento. Foram medidas velocidades de propagação do fogo, taxas de diminuição do peso do combustível e temperaturas atingidas. Os ensaios foram gravados em vídeo e por análise da imagem foi possível obter informação sobre a geometria da chama (altura, inclinação e comprimento de chama). Os parâmetros usados foram a velocidade do vento e a humidade do combustível. Sempre que possível, os resultados obtidos são comparados com resultados existentes na literatura.

ABSTRACT

In order to apply in a correct way the existing models of fire propagation to the portuguese conditions, information is needed on the burning characteristics of the forest fuels most common in Portugal, to allow the fuel sub-model to be adapted (if necessary). This is the main objective of a project under way at the Mechanical Engineering Department, IST.

The present paper describes the experimental set and instrumentation used and shows some results of the work. Beds of pine needles were burned, both in crosswind and in still air. Fire propagation velocities, fuel weight loss rates and temperatures were measured. The experiments were video recorded and, from image analysis, information on flame geometry (flame height, angle with the vertical and flame length) was obtained. Input parameters were wind speed and fuel moisture. Whenever possible, results are compared with data in the literature.

1 - INTRODUÇÃO

A área florestal total do país tem actualmente uma extensão de cerca de 3 milhões de hectares, dos quais 1,3 milhões são de pinheiro bravo (cerca de 45%). A exportação de produtos provenientes da floresta representa cerca de 20% do valor global das nossas exportações e calcula-se que a mão-de-obra empregue no sector florestal é da ordem dos 100 000 trabalhadores, isto é, cerca de 2,5% da população activa.

Da área total indicada, cerca de 75% corresponde a propriedade privada. Dado um número total de quase 400 000 explorações, a área média por exploração é reduzida (5,7 ha).

Desde 1974 foram destruidos pelo fogo mais de 300 000 hectares de matas, o que corresponde a cerca de 10% da área florestal total [1].

A problemática dos fogos florestais no nosso país tem de ser colocada no contexto dos números reproduzidos acima. Enquanto em países como os E.U.A. a gestão das florestas prevê zonas onde fogos que tenham início em causas naturais são deixados arder livremente (desde sempre as florestas viveram em equilíbrio com os incêndios naturais), em Portugal, dada a escassez dos nossos recursos, tal atitude seria impensável[1].

Dois dos parâmetros mais importantes para caracterizar um fogo florestal, com relevância quer nas operações de combate ao fogo, quer nas consequências do fogo (destruição directa e consequências ecológicas a mais longo prazo) são a velocidade de propagação e a intensidade do fogo. Têm sido desenvolvidos vários modelos para descrever o comportamento do fogo e em particular prever, em condições dadas, os valores daqueles dois parâmetros.

Sendo o fogo, e em particular o fogo florestal, um fenómeno extremamente complexo, quer na sua

Nota: O presente artigo constitui uma versão revista da comunicação com o mesmo título apresentada às Jornadas Científicas sobre Incêndios Florestais, realizadas em Coimbra, de 23 a 25 de Novembro de 1988.

Original recebido para publicação em 1 de Julho de 1989

[1] Mesmo nos E.U.A., existe alguma controvérsia neste campo, nomeadamente quando a extensão ardida é grande. Veja-se a discussão havida em relação ao incêndio que no Verão de 1988 cobriu cerca de 400 000 hectares do Parque Nacional de Yellowstone.

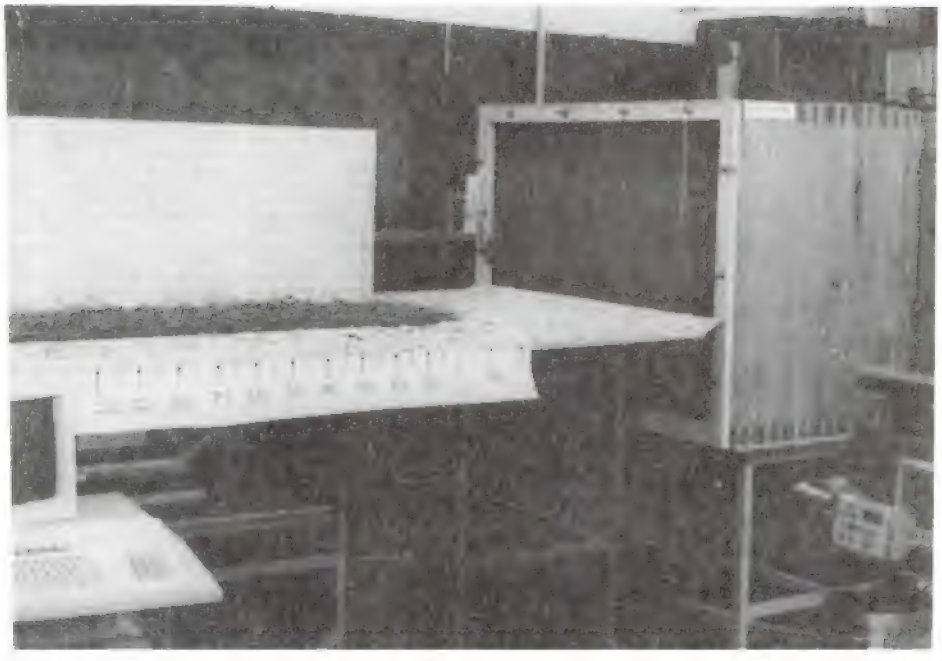


figura 1 - Vista geral da instalação utilizada.

estrutura físico-química, quer na forma como é influenciado pelas condições ambientais, não é de estranhar que os modelos existentes simplifiquem certas características do fogo, afim de tornar o problema tratável. Fons [2], Hottel [3], Anderson [4], Frandsen [5] e Rothermel [6], [7] são alguns dos autores cujos modelos se podem encontrar na literatura. De todos estes, é provavelmente o modelo de Rothermel o mais conhecido; indo buscar a sua expressão para a velocidade de propagação ao modelo de Frandsen, transforma-a depois introduzindo novos termos que, baseados em resultados experimentais, vem a possibilitar a solução aproximada da referida equação. Enquanto o modelo de Frandsen é um modelo analítico, o modelo de Rothermel é semi-empírico; ele tem sido no entanto amplamente testado nos E.U.A., sendo o modelo incorporado no programa BEHAVE, documentado em Andrews [8] e Burgan e Rothermel [9].

2 - INSTALAÇÃO EXPERIMENTAL E INSTRUMENTAÇÃO UTILIZADA

A figura 1 mostra a instalação experimental utilizada no presente trabalho. Colocada à saída da câmara de estabilização de um túnel aerodinâmico (1,20m x 0,90m) existe uma mesa (2,66m x 1,32m) com uma abertura rectangular na qual está posicionado um tabuleiro de alumínio (1,40m x 0,5m) que, por meio de uma estrutura recticulada, assenta sobre o prato de uma balança com saída digital.

Sobre a mesa dispuseram-se alguns obstáculos (pequenas pedras) destinadas a simular a rugosidade existente no terreno em condições reais. No tabuleiro é espalhada uma camada de areia sobre a qual é preparado o leito de combustível. O combustível é inflamado na extremidade do tabuleiro mais próxima da saída do túnel (uma pequena calha com álcool serve de fonte de ignição) e é observada a propagação do fogo ao longo do tabuleiro.

A balança (Mettler PE11) mede o peso do tabuleiro ao longo do tempo e o sinal produzido, à taxa de 4 valores por segundo, é adquirido por um computador HP Vectra, onde fica armazenado para posterior processamento. Ao longo do tabuleiro, em pontos pré-determinados, são colocados conjuntos de três termopares (Pt/Pt - 13% Rh), a alturas diferentes, para permitir obter temperaturas na camada de areia, no leito de combustível e imediatamente acima deste. A figura 2 mostra o desenho esquemático de um desses conjuntos.

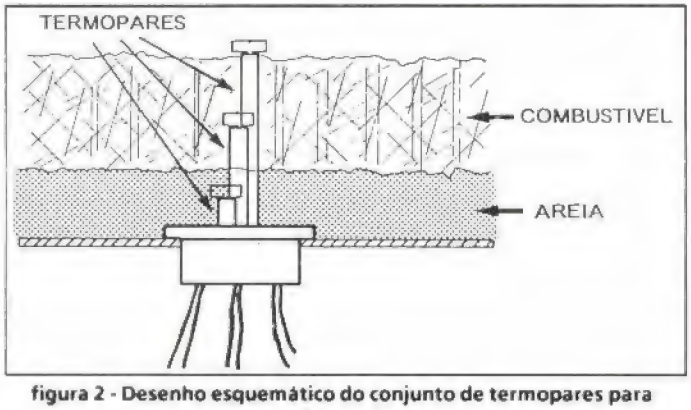


figura 2 - Desenho esquemático do conjunto de termopares para medição de temperaturas.

Posicionada lateralmente em relação ao tabuleiro (canto inferior direito na figura 1) é colocada uma câmara de vídeo (Canon VM-E1), que permite registar a propagação do fogo do início ao fim do tabuleiro, inserindo na imagem o registo do tempo medido por um cronómetro digital. Filmando a propagação do fogo contra um painel quadriculado, é possível obter, por análise posterior da gravação num monitor, informação detalhada sobre a geometria da chama (altura, comprimento e inclinação da chama) e velocidade de propagação.

Foram feitas experiências com diversos tipos de agulhas de pinheiro: *Pinus pinea*, *Pinus halepensis* e *Pinus canariensis*. O presente artigo apenas contém

resultados obtidos com agulhas desta última espécie (pinheiro das Canárias).

3 - RESULTADOS E DISCUSSÃO

Todos os ensaios aqui relatados foram conduzidos para uma mesma carga de combustível de 550 g/m², sendo a altura aproximada do leito de combustível 5 cm e para três valores diferentes da humidade do combustível (peso de água sobre peso de combustível seco): 0%, 7,9% e 18,1%.

A figura 3 apresenta resultados típicos da variação do peso do tabuleiro em função do tempo. Identificada a zona linear da curva, o respectivo declive dá a taxa de perda de peso que nos presentes ensaios, com velocidades do vento relativamente baixas, corresponde essencialmente à taxa de libertação de voláteis, que por sua vez é proporcional à potência calorífica libertada pelo fogo. Para velocidades do vento muito maiores, os resultados da perda de peso virão falseados, pois vai ocorrer o arrastamento de partículas de cinza (e eventualmente de partículas de combustível mais leves) pelo vento.

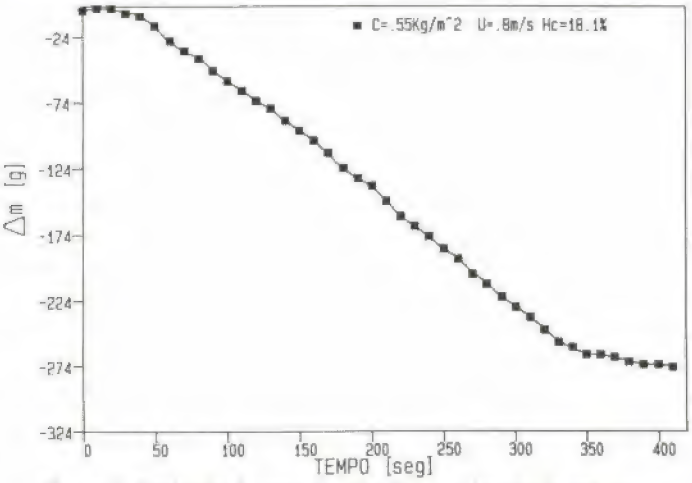


figura 3 - Variação do peso do tabuleiro em função do tempo.

Para a medição de temperaturas, um registador (YEW modelo 3087) recebia o sinal proveniente dos termopares, sendo o intervalo de tempo entre duas temperaturas medidas pelo mesmo termopar de 5 segundos. A figura 4 mostra um exemplo dessas medições. Para maior clareza, o gráfico apenas apresenta as evoluções de temperatura no primeiro grupo de termopares (as evoluções nas outras duas estações de medição são muito semelhantes). Pode observar-se que o termopar 3 -- acima da superfície do leito -- é o que atinge uma temperatura mais elevada, uma vez que fica em contacto directo com a chama, e é também aquele cuja temperatura aumenta em primeiro lugar, pois recebe calor da frente de chama por convecção e radiação. Além disso é o primeiro a diminuir de temperatura, visto que após a passagem da chama fica imediatamente a ser arrefecido por convecção.

A temperatura registada pelo termopar 2 -- dentro do leito de combustível -- aumenta menos e esse aumento tem início mais tarde, porque ele só pode receber calor da chama por condução através do leito (poderá haver algum calor transmitido por convecção e radiação, mas em quantidade muito menor).

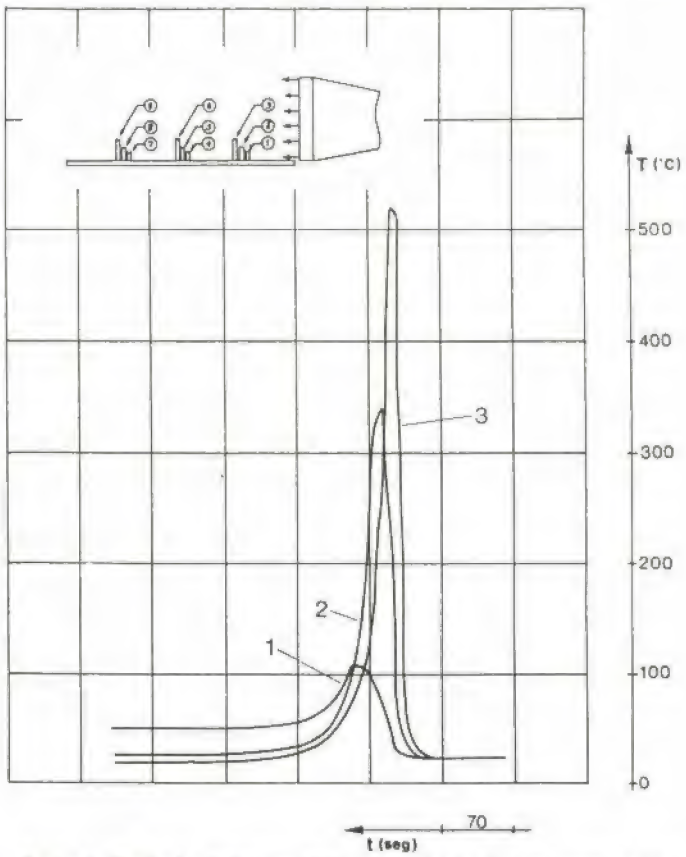


figura 4 - Evolução da temperatura em cada um dos termopares da primeira estação de medição.

Finalmente o termopar 1, inserido na camada de areia, só pode receber calor por condução a partir da superfície; a onda de temperatura atinge-o portanto em último lugar e o valor atingido é muito mais baixo.

Note-se também a diferença nas temperaturas registadas após a passagem do fogo: o termopar 3 regressa rapidamente à temperatura ambiente, o termopar 2 fica a uma temperatura um pouco acima, que é a temperatura das cinzas e que irá diminuindo à medida que estas forem arrefecendo. O termopar 3 é o que fica a uma temperatura mais elevada; a areia arrefece mais lentamente do que o leito de combustível.

Na figura 5 pode ver-se a taxa de variação de massa em função da velocidade do vento, usando como parâmetro a humidade do combustível. Verifica-se que para a mesma velocidade do vento, a taxa de perda de massa diminui com o aumento de humidade. Para uma dada humidade, *m* aumenta com a velocidade do vento, embora os resultados sugiram que esse aumento é menos acentuado quando a humidade do combustível é maior.

A influência da humidade na velocidade de propagação da chama pode observar-se na figura 6. Verifica-se que a velocidade de propagação diminui com o aumento de humidade do combustível, e que para a mesma humidade, a velocidade de propagação aumenta com a velocidade do vento. O carácter basicamente linear da relação encontra-se já documentado, por exemplo em [10] e [11], embora como é sugerido na figura 7 de [4], as linhas tenham que vir a afastar-se da

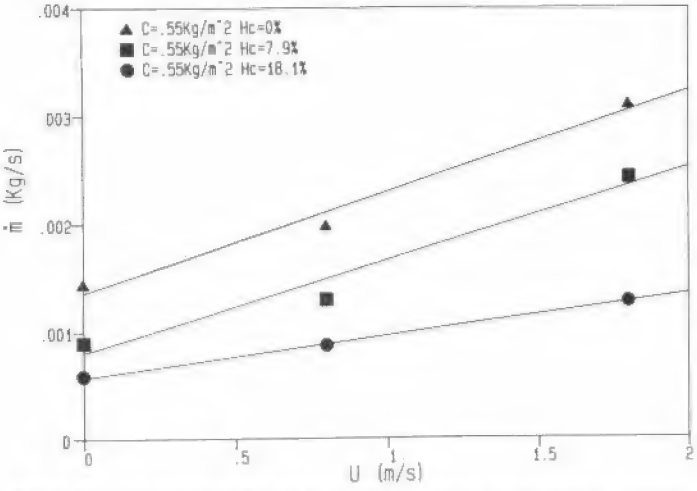


figura 5 - Taxa de variação de peso em função da velocidade do vento para três valores da humidade do combustível.

linearidade para encontrar o eixo das abcissas no valor da humidade de extinção.

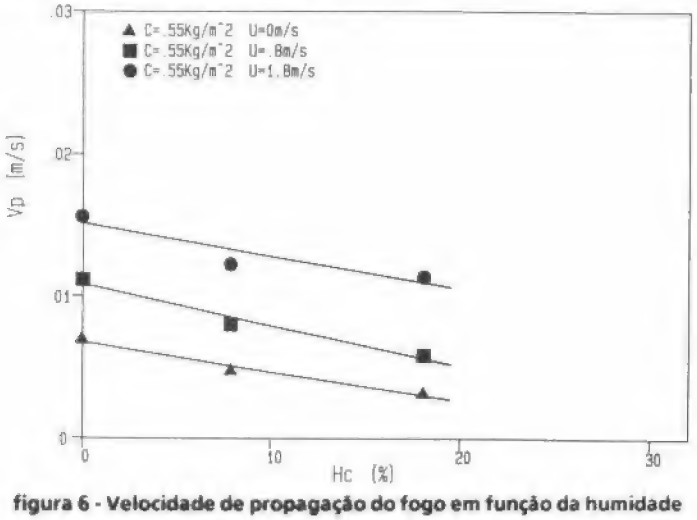


figura 6 - Velocidade de propagação do fogo em função da humidade do combustível, para três valores da velocidade do vento.

A relação entre a velocidade de propagação do fogo e a velocidade do vento está representada na figura 7.

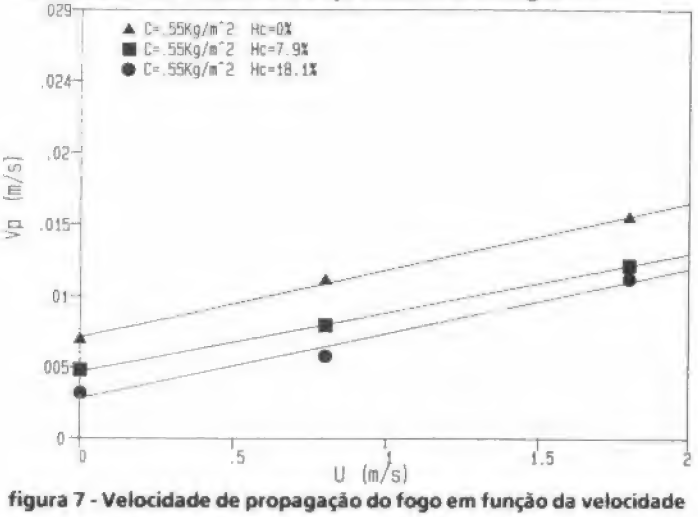


figura 7 - Velocidade de propagação do fogo em função da velocidade do vento, para três valores da humidade do combustível.

A forma destas curvas é discutida em [11], onde se conclui ser ela uma função da razão superfície/volume das partículas de combustível e da porosidade do leito. Para combustíveis finos, como é o presente caso, o declive das curvas aumentaria com a velocidade do vento. A figura 8 apresenta resultados publicados que confirmam esse andamento das curvas; pode ver-se também desta figura que os resultados do presente trabalho estão numa gama de velocidades do vento em que a variação é muito próxima de linear.

O comprimento de chama é um parâmetro com interesse na caracterização do fogo porque pode ser relacionada com a intensidade do fogo (potência calorífica por unidade de comprimento da linha de fogo) através da fórmula de Byram [12]

$$L = 0{,}0775\, I^{0{,}46} \qquad (1)$$

com L em *m* e I em *kW/m*.

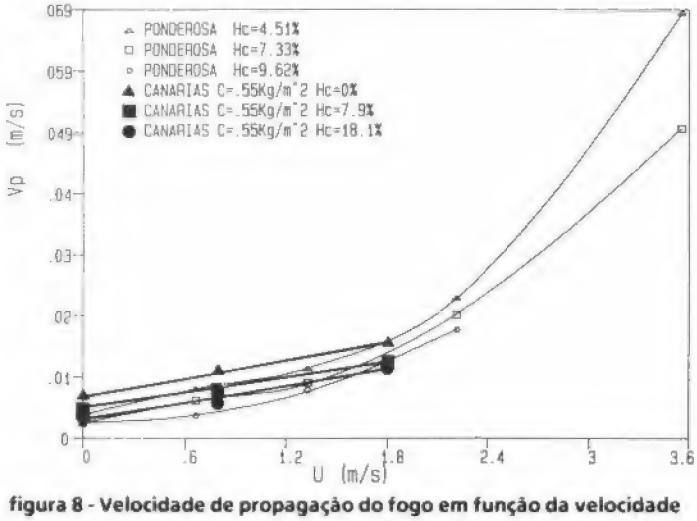


figura 8 - Velocidade de propagação do fogo em função da velocidade do vento. Símbolos cheios: presente trabalho; símbolos abertos: [11].

A figura 9 mostra que o comprimento de chama é praticamente independente da velocidade do vento (pelo menos na gama de velocidades testada); Albini [13] discute razões para esta não dependência, com base na ordem de grandeza das variações da intensidade, velocidade de propagação e velocidade do vento.

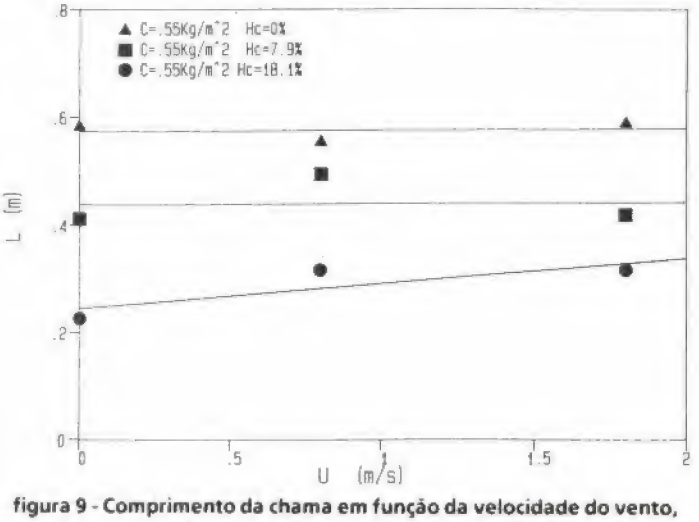


figura 9 - Comprimento da chama em função da velocidade do vento, para três valores da humidade do combustível.

A intensidade do fogo é igual ao produto da taxa de perda de massa (por unidade de comprimento da linha de fogo) pelo poder calorífico do combustível

$$I = m_L P_c \qquad (2)$$

onde m_L vem em kg/(m.s) e P_c em kJ/kg.

Tomando como um valor aproximado para o poder calorífico P_c = 20 000 kJ/kg, e substituindo I na fórmula de Byram, obtem-se

$$L = 7{,}375\, m_L^{0{,}46} \qquad (3)$$

expressão que nos dá o comprimento de chama em função da taxa de perda de massa por unidade de comprimento da linha de fogo.

Na figura 10 pode observar-se a curva representativa da equação (3) em conjunto com o resultado dos presentes ensaios, bem como outros valores tirados da literatura. Não obstante uma certa dispersão dos resultados experimentais, a equação (3) descreve de forma satisfatória a tendência da evolução.

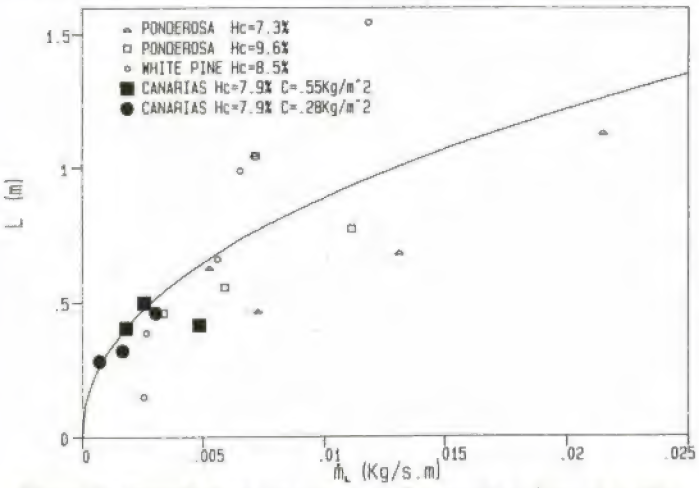


figura 10 - Comprimento da chama em função de mL (taxa de perda de peso por unidade de comprimento da linha de fogo).

Um parâmetro muito utilizado na interpretação de resultados experimentais em fogos é o número de Froude. Definido através de

$$Fr = \frac{U^2}{gH} \qquad (4)$$

onde U é a velocidade do vento, g a aceleração da gravidade e H a altura da chama, o número de Froude pode ser interpretado como a razão entre a força de inércia do vento e a força de impulsão devida à chama, resultante da diferença de temperatura (e portanto de massa específica) em relação ao ar circundante.

A figura 11 mostra a variação da velocidade de propagação com o nº de Froude, tendo como parâmetro a humidade do combustível. Parece haver uma tendência para a diminuição da influência do nº de Froude à medida que este aumenta, mas será necessário efectuar ensaios a nº de Froude mais elevados para o confirmar.

Uma variável importante na propagação do fogo é o ângulo que a chama faz com a vertical; quanto maior for, maior é a quantidade de calor que incide por radiação sobre o combustível não queimado, secando-o e fazendo aumentar a sua temperatura. Rothermel e Anderson [11] representam graficamente a tangente desse ângulo em função de uma variável que definem como o quociente entre a energia do vento (por unidade de área normal à direcção do escoamento e por unidade de

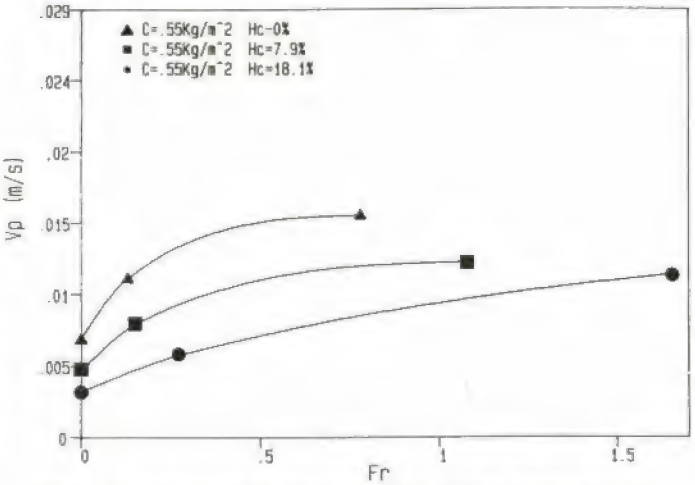


figura 11 - Velocidade de propagação do fogo em função do n de Froude, para três valores da humidade do combustivel.

tempo) e a energia libertada pelo fogo, por unidade de área da zona em combustão e por unidade de tempo. Usando escalas logarítmicas, a relação entre estas duas variáveis é uma recta. Como a variável utilizada por aqueles autores apresenta uma certa analogia com o nº de Froude (uma e outro são uma medida da importância relativa da força do vento e da força de impulsão devida à chama), representou-se na figura 12 a tangente do ângulo θ (ângulo com a vertical) em função do nº de Froude.

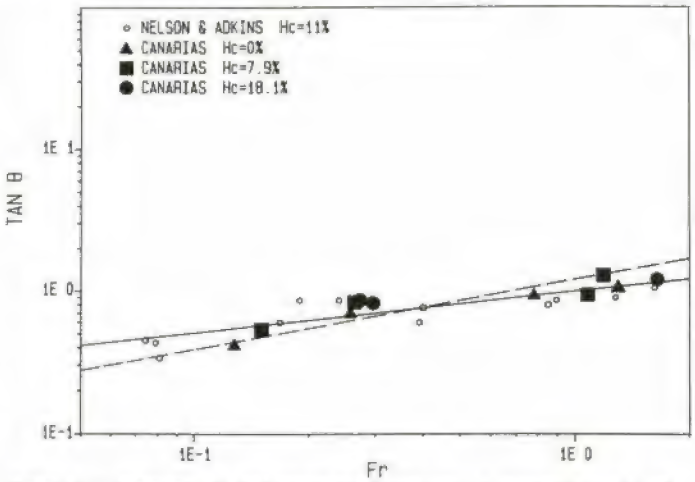


figura 12 - Tangente de θ (ângulo da chama com a vertical) em função do n de Froude. Símbolos cheios: presente trabalho; símbolos abertos: [14]. Linha a traço interrompido: Eq. (5); linha a cheio: Eq. (6).

Os pontos representam medidas do presente trabalho e ainda de Nelson e Adkins [14]; a recta sobreposta a traço interrompido tem a equação

$$\tan\theta = 1{,}22 \left[\frac{U^2}{gH} \right]^{0{,}5} \qquad (5)$$

deduzida por Albini [13], onde U é a velocidade do vento, g a aceleração da gravidade (9,8 m/s^2) e H a altura de chama. Pode observar-se que esta recta não se ajusta de forma aceitável aos pontos representados. A recta representada a traço contínuo é resultante de um ajustamento por mínimos quadrados aos pontos representados. A equação é

$$\tan\theta = 0{,}999\, Fr^{0{,}296} \qquad (6)$$

Nelson e Adkins [14] concluem também que a Eq. (5) não se ajusta aos seus resultados e chegam à equação

$$\tan \theta = 0{,}89\, Fr^{0{,}29} \qquad (7)$$

muito semelhante, como se verifica, à Eq. (6).

4. CONCLUSÕES

Foram apresentados resultados da combustão de resíduos florestais, utilizando parâmetros habituais na caracterização deste tipo de fogos. Foram efectuadas medições da velocidade de propagação do fogo, da taxa de perda de peso e das características geométricas da chama (comprimento, altura e inclinação). Onde apropriado, os resultados obtidos foram comparados com resultados existentes na literatura.

O presente trabalho irá ser continuado ao longo das seguintes linhas:

a) Alargamento da gama dos parâmetros testados, nomeadamente outros valores da humidade, da velocidade do vento e da carga de combustível;
b) repetição dos ensaios com outros combustíveis;
c) utilização de um novo sistema de aquisição dos valores da temperatura;
d) automatização do processo de medição das características geométricas da chama, recorrendo a técnicas automáticas de análise de imagem aplicadas às gravações vídeo dos ensaios.

5. AGRADECIMENTOS

Agradece-se ao Eng. Souto Cruz, do Parque Florestal de Monsanto, as facilidades concedidas para colheita de combustível, bem como todas as informações prestadas sobre combustíveis florestais. Ao Prof. Francisco Rego e Eng. Herminio Botelho, do Departamento Florestal da UTAD, Vila Real, a colaboração prestada ao longo deste trabalho. Partes deste trabalho foram financiadas pela JNICT (Contrato 892.86.152) e DGQA (Projecto I/D 17/87). Agradece-se também à JNICT a concessão de uma bolsa a um dos autores (E.C.F.).

6 - REFERÊNCIAS BIBLIOGRÁFICAS

[1] Wolfgango de Macedo, F. e Sardinha, A. M. (1987). *Fogos Florestais*, 1, Publicações Ciência e Vida, Lda., Lisboa.

[2] Fons, W. (1946), Analysis of fire spread in light forest fuels, *J. Agr. Res.*, **72**, 3, 93-121.

[3] Hottel, H.C., G.C. Williams, G.C. e Steward, F.R. (1965). Modelling of fire spread through a fuel bed, *Tenth Symposium (International) on Combustion*, The Combustion Institute, Pittsburgh, 997-1007.

[4] Anderson, Hal E. (1969). Heat transfer and fire spread, *USDA Forest Service Research Paper INT-69*, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.

[5] Frandsen, William H. (1971). Fire Spread through Porous Fuels from the Conservation of Energy, *Combustion & Flame*, 16, 9-16.

[6] Rothermel, Richard C. (1972). A mathematical model for predicting fire spread in wildland fuels, *USDA Forest Service Research Paper INT-115*, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.

[7] Rothermel, Richard C. (1983). How to Predict the Spread and Intensity of Forest and Range Fires, *USDA Forest Service General Technical Report INT-143*, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.

[8] Andrews, Patricia L. (1986). BEHAVE: Fire Behaviour Prediction and Fuel Modeling System -- BURN Subsystem, Part 1, *USDA Forest Service General Technical Report INT-194*, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.

[9] Burgan, Robert E. e Richard C. Rothermel (1984). BEHAVE: Fire Behaviour Prediction and Fuel Modeling System -- FUEL Subsystem, *USDA Forest Service General Technical Report INT-167*, Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.

[10] Anderson, Hal E. (1964). Mechanisms of fire spread research progress report n.1, *U.S. Forest Service Research Paper INT-8*.

[11] Rothermel, Richard C. e Hal E. Anderson (1966). Fire spread characteristics determined in the laboratory, *U.S. Forest Service Research Paper INT-30*.

[12] Byram, George M. (1959). Combustion of forest fuels, in *Forest fire control and use*, Kenneth P. Davis, ed. McGraw-Hill, New York.

[13] Albini, Frank A. (1981). A Model for the Wind-Blown Flame from a Line Fire, *Combustion & Flame*, 43, 155-174.

[14] Nelson, Ralph M., Jr. e Carl W. Adkins (1986). Flame characteristics of wind driven surface fires, *Can. J. For. Res.*, 16, 1293-1300.

Soluções de um Problema Indeterminado: A Inversão de Dados Sísmicos

João F. B. Duarte Fonseca

Assistente, Departamento de Física do IST.

RESUMO

O problema da inversão sísmica (obtenção de modelos para a crusta e o manto superior a partir de observações à superficie) é apresentado, e os fundamentos da teoria de Backus e Gilbert são discutidos, dando especial relevo aos conceitos de não-unicidade das soluções, resolução espacial e *trade-off* entre resolução e erro. São apresentados alguns detalhes sobre a inversão de máxima verosimilhança, a inversão estocástica e a inversão por decomposição espectral (análise de valores próprios). O método da inversão generalizada é estudado com algum pormenor. Finalmente, um conjunto de dados telessísmicos registados na Estremadura é invertido com o método estocástico [8].

ABSTRACT

The inverse problem of seismology, which consists of modelling the crust and upper-mantle from surface data, is presented. The fundamentals of the theory of Backus and Gilbert are discussed, with emphasis on the concepts of non-uniqueness of solutions, space resolution and error-resolution *trade-off*. Some details are given on maximum-likelywood inversion, stochastic inversion and spectral-decomposition inversion. The generalized inversion is discussed with particular detail. Finally, a set of teleseismic traveltime data recorded in the Portuguese Estremadura is inverted with the stochastic method [8].

1 -Introdução

O objectivo fundamental da sismologia experimental consiste em obter informação sobre a natureza geológica da crusta, a partir de dados obtidos à superficie ou (em menor escala) segundo perfis verticais. Regra geral, pretende-se conhecer a distribuição espacial da velocidade de propagação das ondas sísmicas, a localização das interfaces entre camadas de natureza geológica diversa e a atitude dessas interfaces. A estratificação que normalmente ocorre na crusta terrestre é o ponto de partida para um grande número de técnicas de prospecção, em que se admite que as várias camadas são homogéneas e separadas por interfaces planas, horizontais ou ligeiramente inclinadas. Os métodos mais antigos de prospecção sísmica assentam de facto na hipótese de que a crusta é estratificada, e exploram a informação contida em dois tipos de onda: as ondas que sofrem reflexão nas interfaces (ondas reflectidas) e as ondas que, atingindo as interfaces com o ângulo de incidência crítico, são refractadas ao longo da descontinuidade e regressam à superficie (ondas refractadas ou *head-waves*). A primeira técnica constitui o **método da reflexão**, usado pela primeira vez com fins comerciais em 1927, e a segunda constitui o **método da refracção**, o qual, apesar de ter sido suplantado em aplicação e popularidade pelo anterior, se desenvolveu mais cedo, em 1922. O nascimento da sismologia experimental parece estar ligado à necessidade de localizar com precisão as peças de artilharia pesada inimigas, durante a Primeira Guerra Mundial.

Um método menos restrictivo quanto às configurações admissíveis para o modelo de velocidades é designado por **tomografia sísmica**. Numa das variantes deste método, um volume determinado do interior da Terra é dividido em blocos, no interior dos quais se admite que a velocidade é constante. Num conjunto de estações sismográficas dispostas sobre o volume em investigação, registam-se os tempos de chegada de um conjunto de sismos distantes (tele-sismos) , e comparam-se esses valores com os que são previstos por um modelo de velocidade inicial (geralmente, estabelecido a partir de estudos de reflexão ou refracção). Estabelecem-se relações lineares entre os resíduos dos tempos de chegada e as diferenças entre o modelo real e o modelo inicial, e esse sistema de equações é invertido, permitindo aperfeiçoar o modelo inicial e prosseguir iterativamente.

O número e as dimensões dos blocos em que o volume a analisar é discretizado depende da densidade da "malha" que os raios sísmicos definem no interior do volume, e essa densidade depende criticamente do espaçamento entre as estações e da distribuição azimutal dos epicentros. A discretização em blocos reduz drasticamente a sub-determinação inerente à inversão de dados sísmicos: se a velocidade fôr encarada como

Recebido para publicação em Março de 1987.
Revisto em 18 de Janeiro de 1991.

uma função que varia continuamente no espaço de modo arbitrário, nunca poderá ser resolvida a partir de resultados de observações sismológicas, que constituem sempre um conjunto finito de dados. Mesmo após a discretização do modelo, na generalidade dos casos o problema inverso contém ainda um elevado grau de sub-determinação, e os diferentes métodos de inversão conduzem a uma solução particular que é óptima segundo os respectivos critérios.

A não-unicidade das soluções não é específica da inversão sísmica, já que afecta quase todos os problemas de inversão geofísica, nos domínios da gravimetria, do geomagnetismo, da prospecção eléctrica, etc... Por esse motivo, a teoria da inversão de dados geofísicos tem-se desenvolvido recentemente numa formulação abstracta, sendo depois os resultados aplicáveis aos vários problemas particulares. O ponto de referência no que respeita a esta nova atitude em relação à inversão são os trabalho de BACKUS e GILBERT [1,2], em que se atribui a devida importância a questões como a não-unicidade das soluções, a resolução espacial de um conjunto de dados, o reflexo dos erros experimentais sobre as soluções e a sua relação com a resolução espacial. Com base nesta nova filosofia, a década de setenta conheceu um grande número de técnicas de optimização das soluções do problema inverso.

2 - Teoria geral da inversão

A designação de "problema inverso" justifica-se na medida em que, por definição, uma teoria é capaz de prever os resultados de uma experiência, uma vez estipulados certos parâmetros que regulem o sistema em que a experiência decorre. O problema que agora é discutido descreve essa trajectória em sentido contrário, partindo do conjunto dos resultados de uma experiência (geração, propagação e registo de uma ou várias ondas sísmicas) para chegar ao conhecimento dos parâmetros geofísicos que caracterizam a crusta. Para uma discussão mais detalhada da teoria da inversão, ver TARANTOLA [3].

A discussão da teoria geral da inversão pode ser feita numa base determinística, em que se pressupõe conhecer relações exactas entre os parâmetros que definem o modelo e as quantidades obtidas por observação (afectadas ou não de um erro experimental). Em geral, dispõe-se de uma certa informação a priori sobre os valores que podem tomar os parâmetros, e essa informação pode também assumir a forma de restrições definidas de um modo exacto. Alternativamente, todo o conhecimento de que se dispõe, quer sobre a relação entre dados e parâmetros (a teoria), quer ainda sobre os erros experimentais que afectam os primeiros, quer ainda sobre as restrições nos valores possíveis dos segundos, pode ser encarada de uma prespectiva probabilística, mais difusa mas mais geral. Passaremos a adoptar esta óptica para discutir os conceitos fundamentais envolvidos na teoria da inversão. Posteriormente, e com vista à discussão de aplicações práticas, será analisada a prespectiva determinística.

Na acepção mais geral, o termo "parâmetro" aplica-se a qualquer variável física envolvida num dado problema e que tome um certo valor, seja ele conhecido ou incógnito. Em muitos casos, no entanto, é possível explicitar relações do tipo

$$d_i = g_i\,(p_1, p_2, \ldots, p_j, \ldots, p_m) \qquad i = 1, 2, \ldots, n \tag{2.1}$$

entre as quantidades d_i obtidas experimentalmente e as quantidades p_j que modelam o sistema. Em geral, o termo "parâmetro" é reservado para estas últimas (valor certo, mas incógnito). Porque esta distinção não é intrínseca (é sempre possível conceber a situação inversa em que os parâmetros são conhecidos, e as incógnitas são os "dados" d_i, situação designada por problema directo) é possível enquadrar o problema no espaço $(n+m)$-dimensional de todas as combinações possíveis de valores d_i e p_j, a que chamaremos **Espaço Universal**.

Tendo em conta que um resultado experimental vem afectado de um erro, é mais geral tomar como dado não um valor numérico exacto, mas uma densidade de probabilidade definida sobre todos os valores que a variável em causa pode tomar. Em relação aos parâmetros, a informação apriorística, normalmente na forma de condições de positividade, desigualdades entre parâmetros ou limites superiores e/ou inferiores, pode também ser traduzida através de uma densidade de probabilidade definida no conjunto de todos os valores possíveis. Ambas as informações podem ser fundidas numa distribuição de probabilidade conjunta, que traduza a totalidade de informação disponível antes da inversão, através de uma densidade de probabilidade da forma $\rho(\mathbf{d},\mathbf{p})$ definida no Espaço Universal. Sempre que não houver correlação entre as informações sobre os dados e sobre os parâmetros, é possível escrever

$$\rho(\mathbf{d},\mathbf{p}) = \rho_d(\mathbf{d})\rho_p(\mathbf{p}) \tag{2.2}$$

A relação teórica (2.1) entre dados e parâmetros define no Espaço Universal uma hipersuperfície em geral não linear, que será designada por **variedade teórica**. Se se admitir que a teoria traduzida por (2.1) não é exacta, ou seja, que na sua dedução foram introduzidas hipóteses simplificativas, a variedade teórica passa a ter uma natureza mais difusa, e pode ser substituida por uma função de probabilidade $\theta(\mathbf{d},\mathbf{p})$ definida em todo o Espaço Universal e que traduza para cada ponto a probabilidade de que a variedade teórica passe por esse ponto. Se as simplificações da teoria forem justificáveis, é de esperar que $\theta(\mathbf{d},\mathbf{p})$ seja máxima nos pontos que verificam (2.1). Se a teoria fôr exacta, será

$$\theta\,(\mathbf{d},\mathbf{p}) = \delta\{\mathbf{d} - \mathbf{g}(\mathbf{p})\} \tag{2.3}$$

onde $\mathbf{d} - \mathbf{g}(\mathbf{p}) = 0$ é a versão vectorial da equação (2.1), e $\delta\{.\}$ é o delta de Dirac.

O grau de abstracção a que foi levada a discussão do problema inverso permite obter de imediato uma solução geral, em contraste com a complexidade analítica que se enfrenta quando se aplica a teoria a um problema concreto. Define-se a função densidade de probabilidade a posteriori $\sigma\,(\mathbf{d},\mathbf{p})$ através da expressão

$$\sigma\,(\mathbf{d},\mathbf{p}) = \theta\,(\mathbf{d},\mathbf{p})\,\rho_d(\mathbf{d})\,\rho_p(\mathbf{p}) \tag{2.4}$$

ou seja, a informação apriorística é pesada pela probabilidade teórica $\theta(\mathbf{d},\mathbf{p})$. É interessante notar que a

equação (2.4) contem implícita a resolução quer do problema directo quer do problema inverso, os quais, como já haviamos referido, não são intrinsecamente distintos, mas sim facetas diferentes do problema mais geral que consiste em ajustar informação prévia à luz de uma teoria. É esse problema que é resolvido em toda a sua generalidade pela expressão (2.4).

3 - A Teoria de Backus e Gilbert

Já foi referido no parágrafo 1 que o resultado de uma campanha sísmica é sempre um conjunto finito de dados, referidos a uma rede de estações de observação e a um certo número de localizações do ponto de tiro, e que, a menos que se imponham condições restritivas, o modelo de velocidades é uma função com o contradomínio não discreto, na melhor das hipóteses dependendo apenas de duas ou de uma só coordenada espacial. Desta circunstância resulta que a não-unicidade de solução do problema inverso é tão geral que (enquanto se tratar de inversão sísmica) os casos que a ela fugirem podem ser considerados patológicos. Numa extensa classe de problemas de Geofísica, incluindo a inversão sísmica, é possível estabelecer entre os dados e os parâmetros uma equação integral linear da forma

$$d_i = \int G_i(r)\, m(r)\, dr \qquad (3.1)$$

designada por Equação Integral de Fredholm, a qual vem substituir neste parágrafo a relação (2.1) do parágrafo anterior. Aqui, d_i representa um conjunto finito de dados, supostamente livres de erros experimentais numa primeira análise, m(r) representa o modelo, que por simplicidade é considerado dependente de uma única variável espacial r, e $G_i(r)$ é um núcleo de integração que traduz a teoria. Esta equação, estabelecendo um funcional linear[2] entre o modelo e os dados, aponta já para um aspecto importante da teoria da inversão: a linearização. De facto, alguns problemas não lineares (i.e., nos quais não é possível definir relações do tipo acima) são reduzidos ao caso linear recorrendo ao método das perturbações, através de técnicas standard ou específicas. De qualquer modo serão discutidas oportunamente algumas técnicas de tratamento não linear[3].

2) Designa-se por **funcional** uma aplicação injectiva em que o domínio é um espaço de funções e o contradomínio é R. O qualificativo linear aplicado a um funcional em nada difere do seu significado usual.

3) Mais exactamente, a hipótese importante na teoria da inversão não é a linearidade, mas sim a diferenciabilidade à Frechet. Se os dados estão relacionados com o modelo através do funcional $d_i = F_i(m(r))$, diz-se que F_i é Frechet-diferenciável em m(r) se $F_i(m+\delta m) = d_i + \delta d_i$, com $\delta d_i = \int D_i(m;r)\, m(r)\, dr + 0\{\int (\delta m)^2 dr\}$. $D_i(m;r)$ tomam o nome de derivadas parciais em ordem aos parâmetros. Prova-se que um funcional diferenciável à Frechet e linear pode ser sempre escrito na forma (3.1).

O problema da não unicidade da solução da Equação Integral de Fredholm reduz-se ao problema da existência de soluções da equação homogénea. Seja m_o uma solução não trivial da equação homogénea: então, qualquer função da forma $m_p + c\, m_o$, em que m_p é uma solução particular da equação completa e c é um escalar arbitrário, é ainda uma solução da equação completa, o que prova que estas soluções existem em número infinito. Por outras palavras, se o modelo real tem alguma componente no espaço aniquilador do operador integral $\int G_i(r)\{.\}\, dr$, essa componente não pode ser reconstruida a partir dos dados.

O grau de não unicidade do problema inverso pode ser reduzido seleccionando de entre todos os modelos compatíveis com os dados aqueles que são dotados de significado físico, em face da informação apriorística de que se dispõe, a qual já foi caracterizada anteriormente. A teoria de Backus e Gilbert dá resposta à seguinte questão: que fazer com a classe de todas as funções razoáveis, compatíveis com os dados, e geofisicamente significativas? E a filosofia em que assenta é a seguinte: são relevantes apenas as propriedades que forem comuns a todas as soluções da classe. De facto, se a teoria for exacta, essas propriedades pertencem também ao modelo real.

Backus e Gilbert [1] investigaram as propriedades de funcionais com a forma

$$\langle m \rangle_o = \int A(r_o,r)\, m(r)\, dr \qquad (3.2)$$

em que m(r) é o modelo real e o núcleo de amostragem $A(r_o,r)$ é unimodular, ou seja, $\int A(r_o,r)\, dr = 1$. Para além disto, as médias $\langle m \rangle_o$ devem ser calculáveis exclusivamente a partir dos dados, através de uma relação linear da forma

$$\langle m \rangle_o = \sum_i a_i d_i \qquad (3.3)$$

o que permite concluir de imediato a relação

$$A(r_o,r) = \sum_i a_i G_i(r) \qquad (3.4)$$

como se verifica combinando (3.3) com (3.1) e comparando com (3.2). Então, se as constantes a_i forem escolhidas de modo a que o núcleo de amostragem $A(r_o,r)$ se aproxime tanto quanto possível de um δ de Dirac centrado em r_o, a média $\langle m \rangle_o$ será uma boa estimativa do valor real $m(r_o)$, como mostra a equação (3.2). Uma vez que $\langle m \rangle_o$ é calculado directamente a partir dos dados, os resultados obtidos caracterizam toda a classe de soluções aceitáveis, e são por isso tão significativos como os próprios dados. Em geral, a estimativa $\langle m \rangle_o$ do valor $m(r_o)$ é uma média pesada dos valores m(r) com $r \simeq r_o$, e tem uma resolução espacial limitada pela largura do pico do núcleo $A(r_o,r)$ em torno de r_o. Este conceito de resolução espacial é fundamental em toda a teoria de Backus e Gilbert.

Os critérios de *deltaness*, ou seja, de aproximação entre a função $A(r_o,r)$ e $\delta(r-r_o)$, são variados e, se não apresentam grande dificuldade analítica, são normalmente fonte de morosidade no tratamento numérico. Um processo consiste em minimizar o estimador da largura do pico de $A(r_o,r)$ em torno de r_o

dado por

$$s = 12 \int \left\{ A(r_o,r)(r-r_o)^2 \right\} dr \qquad (3.5)$$

fornecendo o valor minimizado uma medida da resolução espacial obtida.

Por detrás deste método está a hipótese de que o modelo real varia com uma certa suavidade, caso contrário a estimativa $<m>_o$ poderia estar arbitrariamente longe do valor $m(r_o)$ a estimar.

Até este ponto temos admitido que os dados não estão contaminados por erros experimentais, o que é pouco realista. Backus e Gilbert [2] chegaram à conclusão importante de que deve haver um compromisso entre a resolução espacial da solução e a sua contaminação pelos erros experimentais, já que um aperfeiçoamento num dos sentidos implica sempre uma deterioração no outro sentido. Admitindo que os dados d_i estão afectados de um erro gaussiano de média nula e variância comum σ_d^2, a equação (3.3) permite concluir, dada a sua linearidade, que a estimativa $<m>_o$ vem também afectada por um erro gaussiano de variância dada por

$$\sigma_m^2 = \sigma_d^2 \sum_i a_i^2 \qquad (3.6)$$

desde que os erros sejam mutuamente não correlacionados. Tanto a resolução espacial, medida pelo estimador (3.5), como a variância σ_m^2 dependem das constantes a_i, e se se considerar um espaço enquadrado

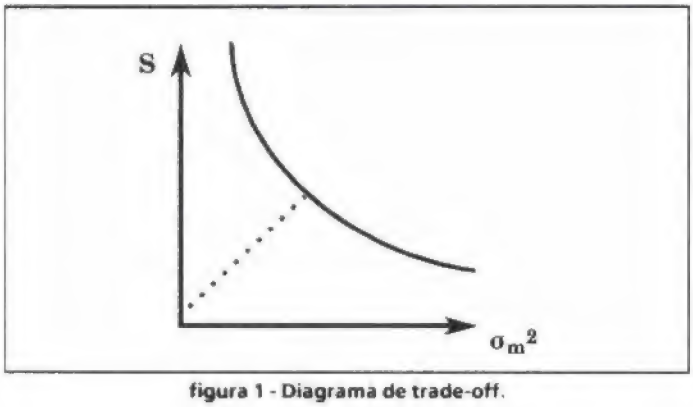


figura 1 - Diagrama de trade-off.

por essas duas quantidades (figura 1) é possivel verificar que, entre todos os conjuntos a_i correspondentes à condição de que $A(r_o,r)$ seja unimodular, o critério de minimização simultânea de ambas as fontes de imprecisão conduz a uma curva do tipo indicada na figura 1. Esta curva toma o nome de diagrama de *trade-off* de resolução e erro. Esse diagrama deve ser explorado em função do compromisso mais conveniente: por exemplo, se se souber de antemão que o modelo real é muito lentamente variável no espaço, será conveniente diminuir a resolução espacial em favor do significado estatístico da solução; se pelo contrário é previsível a ocorrência de saltos bruscos nos valores dos parâmetros, é da precisão que se deve prescindir (afectando os resultados de maior desvio padrão) em favor da resolução espacial.

4 - Algumas soluções particulares do problema inverso

A teoria de Backus e Gilbert não conduz a uma solução do problema inverso, mas sim à maneira de inferir correctamente sobre as propriedades da solução exacta. Outros métodos são distintos, na medida em que apontam para a exploração da variedade teórica à luz de um determinado critério de optimização, conduzindo portanto a uma solução bem definida (que em geral não coincidirá com a solução exacta). Neste parágrafo vamos abordar alguns desses métodos.

Uma via bastante importante que tem sido explorada consiste em tomar como informação apriorística a natureza estatística dos erros que afectam os dados e as estimativas dos modelos. Neste desenvolvimento vamos estudar o caso de modelos discretos, o que permite um cómodo tratamento matricial. Para um problema linear, a teoria será expressa por

$$\mathbf{d} = G\,\mathbf{m} \qquad \mathbf{d} = (d_1, d_2, \ldots, d_n) \qquad (4.1)$$

$$\mathbf{m} = (m_1, m_2, \ldots, m_m)$$

sendo G uma matriz de coeficientes conhecidos. Se fôr possível admitir que os erros nos dados têm uma distribuição gaussiana e que a sua matriz de covariância é C_d (nxn), a sua densidade de probabilidade conjunta será (agora não se impõe que os erros sejam não correlacionados):

$$f(\mathbf{d}) = \frac{|C_d|^{-\frac{1}{2}}}{2\,\pi^{n/2}} \exp\left\{ -\frac{1}{2}(\mathbf{d} - G\,\mathbf{m})^T C_d^{-1} (\mathbf{d} - G\mathbf{m}) \right\} \qquad (4.2)$$

A solução que maximiza esta probabilidade deve minimizar a quantidade

$$(\mathbf{d} - G\mathbf{m})^T C_d^{-1} (\mathbf{d} - G\mathbf{m}) \qquad (4.3)$$

que pode ser interpretada como uma média pesada dos resíduos quadráticos, sendo os pesos os elementos da matriz de covariância. Se os erros forem não correlacionados e tiverem variância comum σ_d^2, será $C_d = \sigma_d^2 I$, e a minimização incide sobre a quantidade $|\mathbf{d}-G\mathbf{m}|^2$ ou seja, cai-se na optimização de minimos quadrados tradicional.

Se se pensar que um problema de inversão passa muitas vezes pela linearização de um problema não linear, utilizando o método das perturbações, é natural que se pretenda minimizar $|\mathbf{m}|^2$, pois isso é particularmente conveniente para o desenvolvimento iterativo em torno de um palpite inicial. No espírito do procedimento anterior, vamos considerar antes a minimização da quantidade

$$\mathbf{m}^T C_m^{-1} \mathbf{m} \qquad (4.4)$$

o que pressupõe algum conhecimento sobre a covariância dos erros que afectam as estimativas dos parâmetros do modelo, através da matriz de covariância C_m. Este conhecimento é em geral de natureza conjectural, e introduz uma subjectividade inerente a este método que é o seu principal ponto fraco.

A técnica que consiste em minimizar conjuntamente as quantidades

$$(\mathbf{d}-G\mathbf{m})^T C_d^{-1}(\mathbf{d}-G\mathbf{m}) \quad \text{e} \quad \mathbf{m}^T C_m^{-1}\mathbf{m} \tag{4.5}$$

toma o nome de **inversão de máxima verosimilhança**. Supostas conhecidas as matrizes C_d e C_m, e tendo em conta que são ambas definidas positivas, é possível escrevê-las na forma

$$C_d = DL_dD^T \qquad C_m = ML_mM^T \tag{4.6}$$

em que as matrizes D e M são compostas pelos vectores próprios de C_d e C_m (que são ortogonais); L_d e L_m são matrizes compostas pelos valores próprios de C_d e C_m na diagonal principal e zeros fora dessa diagonal. O processo de minimização conduz [4] à solução

$$\mathbf{m}_{MV} = ML_m^{\frac{1}{2}} G_g^{-1} L_d^{-\frac{1}{2}} D^T \mathbf{d} \tag{4.7}$$

em que G_g^{-1} é o operador de inversão generalizada que será definido no parágrafo seguinte.

A inversão de máxima verosimilhança passa pelo cálculo do operador de inversão generalizada, o que exige o recurso a um processo de decomposição espectral da solução que nas aplicações com interesse se torna caro em termos de computação. Um procedimento diverso e que requere apenas multiplicações e inversões de matrizes foi desenvolvido sob a designação de **inversão estocástica**. Admite-se que os dados se podem decompôr em sinal e ruido, segundo

$$\mathbf{d} = G\mathbf{m} + \mathbf{n} \tag{4.8}$$

sendo **m** e **n** processos estocásticos, ou seja, famílias de variáveis aleatórias, de média nula e com matrizes de covariância C_{mm} e C_{nn}, respectivamente. Procura-se o operador matricial L, designado operador de inversão estocástica, que minimize o resíduo quadrático $|\mathbf{m} - L\mathbf{d}|^2$.

O processo clássico de minimização conduz à expressão

$$L = C_{md}C_{dd}^{-1} \tag{4.9}$$

sendo C_{dd} a matriz de covariância dos dados e C_{md} a matriz de covariância cruzada entre os dados e o modelo. Impondo a independência estatística entre **m** e **n**, conclui-se por linearidade que

$$C_{dd} = GC_{mm}G^T + C_{nn} \tag{4.10}$$

$$C_{md} = C_{mm}G^T \tag{4.11}$$

e finalmente a solução pode ser escrita na forma

$$\mathbf{m}_s = C_{mm}G^T C_{dd}^{-1}\mathbf{d} \tag{4.12}$$

De passagem, registe-se um caso especialmente importante, que corresponde a admitir-se que as covariâncias são da forma $C_{mm} = \sigma_m^2 I_m$ e $C_{nn} = \sigma_n^2 I_n$; neste caso o operador de inversão estocástica admite a seguinte expressão:

$$\begin{aligned} L_0 &= C_{mm}G^T(GC_{mm}G^T + C_{nn})^{-1} \\ &= G^T(G^TG + \varepsilon^2 I_n)^{-1} \end{aligned} \tag{4.13}$$

$$\text{com} \quad \varepsilon^2 = \frac{\sigma_n^2}{\sigma_m^2}$$

Este algorítmo toma o nome de **operador de mínimos quadrados amortecidos**, por motivos que serão discutidos mais adiante.

Dada a imprecisão intrínseca dos dados experimentais, e em muitos casos a natureza aproximada das teorias, a solução do problema inverso tem geralmente as duas seguintes características: a solução obtida não verifica os dados experimentais, quando a partir dela se resolve o problema directo; e os dados não contêm informação sobre determinadas componentes do modelo real. A análise de vectores próprios matriciais, ou desenvolvimento espectral, método aplicado à resolução de sistemas lineares pela primeira vez por LANCZOS [5], tem a vantagem de deixar perfeitamente claro, ao longo da resolução, quais as zonas do espaço de modelos que não são "iluminadas" por um conjunto de dados, e quais as zonas do espaço de dados cujos pontos não podem ser obtidos resolvendo o problema directo, qualquer que seja o modelo.

Considere-se a equação

$$d_i = \int G_i(r)\, m(r)\, dr \qquad i = 1,2,\ldots,n \tag{4.14}$$

em que admitiremos que os dados estão contaminados de erros gaussianos e ainda que d_i e G_i foram normalizados pelo desvio-padrão σ_i dos dados, de modo a que estes tenham variância unitária. A matriz definida por

$$T_{ij} = \int G_i(r)\, G_j(r)\, dr \tag{4.15}$$

é por construção definida positiva e simétrica, e pode ser diagonalizada na forma

$$M^T T M = L \tag{4.16}$$

sendo a diagonal de L composta pelos valores próprios de T colocados por ordem decrescente:

$$L = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ldots & \\ 0 & & & \lambda_n \end{bmatrix}$$

com $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \geq 0$.

Prova-se que as funções definidas por

$$\psi_i(r) = \lambda_i^{-\frac{1}{2}} \sum M_{ij} G_j(r) \tag{4.17}$$

constituem um sistema ortonormado [6] e que é possível decompôr qualquer modelo na forma

$$m = \sum a_i \psi_i + \psi^* \tag{4.18}$$

sendo a_i, a projecção de m(r) segundo $\psi_i(r)$, dada por

$$a_i = \int \psi_i(r)\, m(r)\, dr = \lambda_i^{-\frac{1}{2}} \sum M_{ij} d_j \tag{4.19}$$

O resíduo ψ^* inclui qualquer componente de m(r) que seja solução da equação homogénea, imprevisível a partir dos dados, e ainda as componentes de m(r) associadas a valores próprios nulos. Um importante resultado é o seguinte: o desvio padrão de que vem afectado cada coeficiente a_i é justamente $\lambda_i^{-1/2}$, o que permite considerar (4.18) como a decomposição de m(r) numa base em que a incerteza é crescente com o índice. Então podem ser convenientemente filtradas as componentes da solução associadas a valores próprios pequenos. Deste modo, a solução ficaria a um tempo mais suavizada e menos afectada pelos erros experimentais, o que é novamente o compromisso a que alude a teoria de Backus e Gilbert.

O número de componentes a considerar no desenvolvimento introduz um factor subjectivo, mas pode ser determinado pelo seguinte critério: quanto mais extensa for a truncatura maior será a discrepância entre dados d_i e os valores d_i' obtidos por solução do problema directo a partir da solução do problema inverso. Então, a quantidade.

$$x^2 = \sum (d_i - d_i')^2 \tag{4.20}$$

pode ser tomada para controlar o ajustamento. Verifica-se que esta estatística tem uma distribuição de χ^2 com (n-p) graus de liberdade, pelo facto de os d_i serem gaussianos, sendo p o número de parcelas tomadas no desenvolvimento.

Os últimos métodos analisados têm como ponto de partida a existência de uma relação linear entre dados e parâmetros do modelo. Já foi visto que esta linearidade é muitas vezes uma aproximação, e a convergência de um tal esquema pode ficar comprometida sempre que a não linearidade do problema for severa. Veremos em seguida alguns métodos para lidar com a não-linearidade.

Seja $\mathbf{d} = \mathbf{g}(\mathbf{m})$ uma relação teórica não linear, e seja $\mathbf{m}_0$ um palpite inicial. O procedimento mais directo seria aplicar o método dos mínimos quadrados, minimizando o resíduo quadrático médio dado por

$$\phi = |\mathbf{d} - \mathbf{g}(\mathbf{m})|^2 \tag{4.21}$$

Se se admitir que

$$\mathbf{g}(\mathbf{m}) \simeq \mathbf{g}(\mathbf{m}_0) + G\mathbf{m}' \quad \text{com} \quad \mathbf{m}' = \mathbf{m} - \mathbf{m}_0$$
$$\text{e} \quad G_{ij} = \left(\frac{\partial g_i}{\partial m_j}\right)_{\mathbf{m}=\mathbf{m}_0} \tag{4.22}$$

pode escrever-se que

$$\phi = |\mathbf{d}' - G\mathbf{m}'|^2 \quad \text{com} \quad \mathbf{d}' = \mathbf{d} - \mathbf{g}(\mathbf{m}_0) \tag{4.23}$$

A solução de mínimos quadrados pode ser escrita na forma standard

$$\mathbf{m}_L' = (G^T G)^{-1} G^T \mathbf{d} \tag{4.24}$$

e o processo iterativo procede definindo $\mathbf{m}_1 = \mathbf{m}_0 + \mathbf{m}_L$ e introduzindo a nova relação linear

$$\mathbf{g}(\mathbf{m}) = \mathbf{g}(\mathbf{m}_1) + G\mathbf{m}'' \quad \text{com} \quad \mathbf{m}'' = \mathbf{m} - \mathbf{m}_1$$
$$\text{e} \quad G_{ij} = \left(\frac{\partial g_i}{\partial m_j}\right)_{\mathbf{m}=\mathbf{m}_1} \tag{4.25}$$

Na maioria das situações com interesse, o sistema (4.24) não pode ser resolvido. Mesmo para os casos em que G^TG é não-singular, este esquema iterativo pode ser divergente, em face de uma não-linearidade forte. O **método do gradiente** é um processo alternativo de convergência garantida, mas que perde em rapidez de convergência, já que é um método extremamente lento. Aqui, a iteração é feita na direcção de máxima variação da quantidade ϕ. O plano tangente à superfície $\phi = \phi_0$ no ponto $\mathbf{m} = \mathbf{m}_0$ tem por equação

$$\Sigma \frac{\partial \phi}{\partial \mu_j}(m_j - m_{0j}) = 0 \tag{4.26}$$

e o vector que lhe é perpendicular (apontando no sentido decrescente de ϕ) tem por componentes as derivadas parciais

$$\left(-\frac{\partial \phi}{\partial m_j}\right)_{m=m_0}.$$

Este vector será designado por $\mathbf{m}'_G$. Tendo em conta a expressão de ϕ, este vector tem componentes

$$m'_{G_j} = -2 \sum_i G_{ij} d_i \tag{4.27}$$

$$G_{ij} = \left(\frac{\partial g_i}{\partial m_j}\right)_{\mathbf{m}=\mathbf{m}_0}.$$

Dada a lenta convergência deste processo iterativo, foi sugerido que a iteração avançasse numa direcção intermédia entre as de $\mathbf{m}'_L$ e $\mathbf{m}'_G$. A equação (4.24) dá a expressão de $\mathbf{m}'_L$, e a equação (4.27) permite concluir que $\mathbf{m}'_G$ tem a direcção de $G^T\mathbf{d}$. A solução intermédia pode ser expressa na forma

$$\mathbf{m}'_{DL} = (G^T G + \varepsilon^2 I)^{-1} G^T \mathbf{d} \tag{4.28}$$

e tem a direcção de $\mathbf{m}'_L$ quando $\varepsilon = 0$, aproximando-se assimptoticamente da direcção de $\mathbf{m}'_G$ quando ε cresce. Comparando (4.28) com (4.13), é interessante verificar que esta estimativa, designada por **solução de mínimos quadrados amortecidos**, não é mais do que a solução estocástica no caso especial então considerado. De notar também que $(G^TG + \varepsilon^2 I)$ é sempre não-singular.

A exploração do espaço de modelos para a obtenção de um palpite inicial numa zona de mínimos da função ϕ pode ser feita através da geração de pontos do espaço de uma forma aleatória, investigação e selecção desses pontos em função do correspondente valor de ϕ. Nisto consiste essencialmente o método de Monte Carlo[4]. Naturalmente, é um método extremamente lento, o que é um preço a pagar quando o problema é não linear. Quando o problema é linear, e se conhecem limites superiores e inferiores para os parâmetros do modelo, essas condições (elas próprias não lineares) definem no

espaço de modelos um poliedro convexo, e a exploração do espaço no sentido da minimização de ϕ pode ser feita apenas sobre os vértices do poliedro, já que a teoria prevê que apenas sobre esses vértices possam ocorrer os valores extremos. A exploração é neste caso extremamente rápida e eficaz, e constitui um exemplo de aplicação de uma importante técnica de resolução de problemas lineares, designada por programação linear.

5 - Inversão generalizada

O **método da inversão generalizada** apresenta uma solução para o problema linear

$$\mathbf{d} = G\mathbf{m} \tag{5.1}$$

em que G é uma matriz NxM de coeficientes conhecidos. A solução é obtida pelo método da expansão espectral já discutido.

O Espaço Universal pode ser dividido em dois subespaços, da forma $(\mathbf{d},\mathbf{0})$ e $(\mathbf{0},\mathbf{m})$, com $\mathbf{d}=(d_1,\ldots,d_N)$ e $\mathbf{m}=(m_1,\ldots,m_M)$, a que chamaremos respectivamente o espaço de dados e o espaço de modelos (esta terminologia, por óbvia, já foi utilizada anteriormente em vários pontos). A matriz definida por

$$T = \left| \begin{array}{c|c} 0 & G \\ \hline G^T & 0 \end{array} \right| \tag{5.2}$$

é hermítica por construção, e por isso tem N + M valores próprios reais, soluções da equação

$$\det(T - \lambda I) = 0 \ . \tag{5.3}$$

O vector próprio associado a cada valor próprio λ_i verificará

$$T\mathbf{w}_i = \lambda_i \mathbf{w}_i \tag{5.4}$$

e, se se considerar $\mathbf{w}_i$ como a soma de dois vectores, um contido no espaço de dados e outro contido no espaço de modelos, de acordo com

$$\mathbf{w}_i = (\mathbf{u}_i \,|\, \mathbf{0}) + (\mathbf{0} \,|\, \mathbf{v}_i)$$

a forma especial da matriz T permite escrever

$$\begin{aligned} G\,\mathbf{v}_i &= \lambda_i \mathbf{u}_i \\ G^T \mathbf{u}_i &= \lambda_i \mathbf{v}_i \end{aligned} \tag{5.5}$$

Este importante resultado estabelece que uma teoria linear, aqui representada pela matriz G, estabelece a ligação entre pontos do espaço de dados e pontos do espaço de parâmetros (i.e., permite inverter os dados) apenas através dos valores próprios não nulos da matriz T. Para os valores próprios nulos será

$$\begin{aligned} G\mathbf{v}_i &= 0 \\ G^T \mathbf{u}_i &= 0 \ . \end{aligned} \tag{5.6}$$

As equações (5.5) permitem escrever que

$$\begin{aligned} G^T G \mathbf{v}_i &= \lambda_i^2 \mathbf{v}_i \\ G G^T \mathbf{u}_i &= \lambda_i^2 \mathbf{u}_i \end{aligned} \tag{5.7}$$

e, tendo em conta que as matrizes G^TG e GG^T são hermíticas, conclui-se que os vectores $\mathbf{u}_i$ constituem uma base ortogonal completa para o espaço de parâmetros. Se esses vectores, uma vez calculados, forem normalizados, é possível escrever

$$\begin{aligned} \mathbf{v}_i^T . \mathbf{v}_j &= \delta_{ij} \qquad i,j = 1,2,\ldots,M \\ \mathbf{u}_i^T . \mathbf{u}_j &= \delta_{ij} \qquad i,j = 1,2,\ldots,N \end{aligned} \tag{5.8}$$

Verifica-se facilmente que se λ_i é solução da equação aos valores próprios (5.3) associada ao vector próprio $(\mathbf{u}_i \,|\, \mathbf{v}_i)$, então $-\lambda_i$ também é solução, associada ao vector próprio $(-\mathbf{u}_i \,|\, \mathbf{v}_i)$. Seja 2p o número de valores próprios $\pm\lambda_i$ diferentes de zero: existirão N + M-2p valores próprios nulos. As equações aos valores próprios (5.7) terão p soluções não nulas cada, o que permite decompôr a matriz U (NxN) cujas colunas são formadas pelos vectores próprios $\mathbf{u}_i$ na forma

$$U = (U_p \,|\, U_0) \tag{5.9}$$

em que U_p incorpora os vectores próprios associados a valores próprios não nulos. De modo semelhante se pode construir a matriz V (MxM) cujas colunas são os vectores próprios $\mathbf{v}_i$, e decompô-la na forma

$$V = (V_p \,|\, V_0) \tag{5.10}$$

Nesta nova linguagem matricial, as equações (5.5) são

$$\begin{aligned} G V_p &= U_p L_p \\ G^T U_p &= V_p L_p \end{aligned} \tag{5.11}$$

sendo L_p uma matriz diagonal formada pelos valores próprios λ_i não nulos. Esta matriz fora já referida no parágrafo anterior, quando pela primeira vez se abordou a temática do desenvolvimento espectral (por espectro do problema entende-se o conjunto dos valores próprios λ_i, e daí a designação). Para as matrizes U_0 e V_0 valerá

$$\begin{aligned} G V_0 &= 0 \\ G^T U_0 &= 0 \end{aligned} \tag{5.12}$$

As quatro últimas equações matriciais podem ser condensadas na forma

$$GV = G\,(V_p \,|\, V_0) = (U_p \,|\, U_0)\,(L_p \,|\, 0) \tag{5.13}$$

ou ainda

$$G = (U_p \,|\, U_0)\,(L_p \,|\, 0)\,(V_p \,|\, 0)^T = U_p L_p V_p^T \tag{5.14}$$

Esta factorização põe em evidência que a teoria traduzida pela matriz G nada nos pode esclarecer

4) Outros métodos de inversão, normalmente associados aos nomes de CHERNOV e de KEILIS-BOROK, utilizam a exploração aleatória exaustiva do espaço de modelos como meio de realizar a própria inversão. Nisso são distintos da aplicação agora referida, que apenas conduz a um palpite inicial.

acerca das componentes dos parâmetros que pertençam ao espaço V_0, nem pode prever componentes dos dados que pertençam ao espaço U_0. Por espaço V_0 entenda-se o espaço gerado pelas combinações lineares das colunas de V_0, etc...

Qualquer vector de V_0 adicionado a uma dada solução não afecta os dados (obtidos por resolução do problema directo), e esta é a essência da não-unicidade do problema inverso.

Quando a matriz ULV^T possui uma inversa (problema determinado) ela pode ser escrita na forma $G^{-1} = VL^{-1}U^T$, já que $UU^T = U^TU = I_N$ e $VV^T = V^TV = I_N$, como se conclui das relações (5.8). Por analogia, a inversa generalizada de G é definida por

$$G_g^{-1} = V_p L_p^{-1} U_p^T \tag{5.15}$$

A solução inversa generalizada é definida em conformidade por

$$\mathbf{m}_g = G_g^{-1}\mathbf{d} \ . \tag{5.16}$$

É importante notar que num problema determinado (i.e., tantas equações independentes como incógnitas) a matriz G é não-singular, e por isso não tem valores próprios nulos. Nesse caso a matriz admite uma inversa exacta, que coincide com a inversa generalizada, pois as matrizes U_p e V_p são "completas". Que a inversa generalizada degenere na inversa exacta quando o problema é determinado é um resultado que seria de esperar, pois em caso contrário dificilmente se poderia atribuir qualquer significado àquela solução.

Quando todos os valores próprios nulos estão associados a vectores próprios do espaço de dados (i.e., existe espaço U_0 mas não existe espaço V_0) a solução generalizada verifica

$$\mathbf{m}_g = (G^T G)^{-1} G^T \mathbf{d}$$

ou

$$G^T G\, \mathbf{m}_g = G^T \mathbf{d} \tag{5.17}$$

o que mostra que nas presentes condições a inversa generalizada degenera na solução de mínimos quadrados standard, pelo que se pode concluir que ela minimiza a norma $|\mathbf{d} - G\,\mathbf{m}_g|$.

Quando, pelo contrário, o espaço U_0 é nulo mas existe V_0 (problema com solução não-única mas em que os dados podem ser calculados de modo exacto a partir de qualquer das várias soluções) a solução generalizada verifica os dados, como veremos em seguida:

$$G\mathbf{m}_g = G\,G_g^{-1}\mathbf{d} = U_p L_p V_p^T V_p L_p^{-1} U_p^T \mathbf{d} \tag{5.18}$$

Ora a igualdade $V_p^T V_p = I_p$ verifica-se sempre (bem como $U_p^T U_p = I_p$), dada a ortonormalidade dos vectores próprios $\mathbf{v}_i$. Porém, a igualdade $V_p V_p^T = I_p$ (ou $U_p U_p^T = I_p$) só se verifica para matrizes quadradas. No presente caso,a matriz U_p é NxN, e por isso verifica-se $U_p^T U_p = I_N$, o mesmo não se podendo afirmar do produto $U_p U_p^T$. Nestas condições, a equação (5.18) permite concluir que

$$G\,\mathbf{m}_g = \mathbf{d} \tag{5.19}$$

como pretendiamos verificar.

Se a solução generalizada verifica os dados, qualquer outra solução que os verifique deve diferir de $\mathbf{m}_g$ apenas nas componentes que possuir no espaço V_0. Então será

$$\mathbf{m} = \mathbf{m}_g + \sum a_i \mathbf{v}_i \tag{5.20}$$

e

$$|\mathbf{m}|^2 = |\mathbf{m}_g|^2 + \sum a_i^2 \quad (\text{dado que } \mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}) \tag{5.21}$$

Esta equação permite concluir que, quando não existe espaço U_0, a inversão generalizada dá a solução com menor norma de entre todas as que verifiquem os dados.

No caso geral, em que existem ambos U_0 e V_0, a inversão generalizada fornece uma solução que minimiza simultaneamente $|\mathbf{d} - G\mathbf{m}|$ e $|\mathbf{m}|$.

O formalismo matricial da inversão generalizada permite analisar comodamente a resolução e o erro da solução. De facto, tendo em conta que

$$\mathbf{m}_g = G_g^{-1}\mathbf{d} = (G_g^{-1} G)\,\mathbf{m} \tag{5.22}$$

conclui-se que a solução $m_{gi} = m_g(r_i)$ é uma média pesada dos valores $m_g(r_j)$ em torno de r_i, sendo os pesos os elementos da matriz $G_g^{-1}G$, que toma o papel do núcleo de amostragem da teoria de Backus e Gilbert. Esta matriz indica a resolução espacial da solução, a qual será óptima quando $G_g^{-1}G = I$. Tendo em conta a equação (5.15), será

$$G_g^{-1} G = V_p L_p^{-1} U_p^T L_p V_p^T = V_p V_p^T \tag{5.23}$$

Novamente se conclui que a resolução é ideal quando não há valores próprios nulos associados ao espaço de modelos. Nessas condições será $V_p V_p^T = I_M$, e a solução é única, ainda que os dados sejam imperfeitos.

Por outro lado, se for $\mathbf{\Delta d}$ o erro que afecta os dados experimentais, o erro resultante na solução inversa generalizada será (dada a linearidade do problema)

$$\Delta\mathbf{m}_g = G_g^{-1}\Delta\mathbf{d} \tag{5.24}$$

A matriz de covariância dos erros $\mathbf{\Delta m}_g$ será

$$C_{mm} = \left\{ E(\Delta m_{gi}\, \Delta m_{gj}) \right\} = G_g^{-1} C_{dd} (G_g^{-1})^T$$

sendo C_{dd} a matriz de covariância dos erros nos dados. Admitindo que estes últimos são gaussianos, estatisticamente independentes, e que têm a mesma variância σ^2, será

$$C_{mm} = \sigma^2 G_g^{-1} (G_g^{-1})^T = \sigma^2 V_p L_p^{-1} U_p^T U_p L_p^{-1} \tag{5.25}$$
$$= \sigma^2 V_p L_p^{-2} V_p^T$$

o que vem confirmar a asserção anterior segundo a qual os valores próprios muito pequenos contribuem para aumentar a variância da solução, já que os elementos da matriz representada por L_p^{-2} são λ_i^{-2}. O número p deve ser escolhido de modo a que, ignorando os valores próprios próximos de zero, se diminua a variância do erro da solução, sem que isso afecte damasiado a resolução espacial. Esta questão de *trade-off* fora já discutida aquando do estudo da teoria de Backus e Gilbert.

6 - Modelação tomográfica da litosfera da Estremadura

Apresenta-se em seguida a aplicação à Estremadura de um método desenvolvido por Aki et al.[7]. Para proceder à inversão, a crusta e o manto superior são divididos em camadas horizontais, as quais por sua vez se subdividem em blocos homogéneos, cada um dos quais caracterizado por um valor v_i da velocidade de propagação. A solução é linearizada em torno de um modelo inicial, em que a velocidade é constante dentro de cada camada.

Considere-se um raio que atravessa a camada i, percorrendo no seu interior o espaço d_i, com a velocidade v_i. Seja $\bar{v}_i$ velocidade média da camada e $\bar{d}_i$ o espaço que o raio percorria na camada com o modelo inicial de velocidades. Definindo a perturbação fraccional do inverso da velocidade, m_i, através da expressão

$$\frac{1}{v_i} = \frac{1}{\bar{v}_i}(1+m_i) \tag{6.1}$$

o tempo de percurso real dentro da camada i será

$$t_i = \frac{d_i}{v_i} = \left(\frac{\bar{d}_i + \Delta d_i}{\bar{v}_i}\right)(1+m_i) \tag{6.2}$$

com $\Delta d_i = d_i - \bar{d}_i$.

Se for NL o número de camadas do modelo, o tempo de chegada de um raio que atravessa todas as camadas será

$$t = \sum_i^{NL} t_i + \tau + \varepsilon = \sum_i^{NL} \frac{\bar{d}_i}{\bar{v}_i}(1+m_i) + \tau + \varepsilon =$$

$$= \sum_i^{NL} \frac{\bar{d}_i}{\bar{v}_i}(1+m_i) + \sum_i^{NL}\left(\frac{\Delta d_i}{\bar{v}_i}\right)(1+m_i) + \tau + \varepsilon \tag{6.3}$$

sendo τ o tempo de chegada à base do modelo e ε uma parcela que dá conta dos erros experimentais. Com o modelo inicial o tempo de chegada à base do modelo (calculado a partir dos dados) seria $\bar{\tau} = \tau - \Delta\tau$, e a expressão (6.3) pode ser escrita na forma

$$t = \bar{\tau} + \sum_i \frac{\bar{d}_i}{\bar{v}_i}(1+m_i) + \varepsilon' \tag{6.4}$$

com

$$\varepsilon' = \varepsilon + \Delta\tau + \sum_i^{NL} \frac{\Delta d_i}{v_i}(1+m_i) \tag{6.5}$$

É legítimo associar o erro experimental com as duas últimas parcelas de (6.3) porque, de acordo com o princípio de Fermat, o tempo de percurso é estacionário em relação a perturbações do percurso em torno do raio real, e essas parcelas são de ordem superior no desenvolvimento (6.4).

Ao atravessar a camada i através do bloco j , um raio pode sofrer ligeiras contribuições de blocos vizinhos. Por esse motivo, a perturbação m_{ij} corresponde ao bloco j da camada i pode ser decomposta na forma

$$m_{ij} = \bar{m}_{ij} + \Delta m_{ij} \tag{6.6}$$

em que m_{ij} é a perturbação média na camada i calculada a partir dos raios que atravessam essa camada quase exclusivamente pelo bloco j, e Δm_{ij} é um termo residual. A perturbação m_i que afecta o raio no interior da camada i poderá ser escrita na forma

$$m_i = \sum_j^{NB} \bar{m}_{ij} F_{ij} + \Delta m_i \tag{6.7}$$

onde F_{ij} é um indicador que toma o valor 1 se o raio atravessa a camada i quase exclusivamente através do bloco j, e é nulo em caso contrário. NB é o número de blocos por cada camada.

Seja $\bar{t}$ o tempo que um dado raio leva a atravessar o modelo inicial, ou seja, $\bar{t} = \Sigma\, \bar{d}_i / \bar{v}_i$. O desenvolvimento (6.4) pode ser reescrito, tendo em conta (6.7), na forma

$$t - \bar{t} = \bar{\tau} + \sum_i \sum_j g_{ij}\, \bar{m}_{ij} + \varepsilon''$$

$$\text{com} \quad g_{ij} = F_{ij} \frac{\bar{d}_i}{\bar{v}_i} \tag{6.8}$$

tendo a parcela ε'' absorvido os resíduos Δm_i.

Se a rede de observações dispõe de um total de N estações, e designando por M = NBxNL o número de blocos de todo o modelo, as N equações correspondentes a (6.8) podem ser condensadas na forma

$$\mathbf{t} - \bar{\mathbf{t}} = \bar{\boldsymbol{\tau}} + G\mathbf{m} + \boldsymbol{\varepsilon}'' \qquad \text{com} \quad G = \{g_{ij}\}_{N\times M} \tag{6.9}$$

As quantidades $\bar{\mathbf{t}}$, $\bar{\boldsymbol{\tau}}$ e $\boldsymbol{\varepsilon}''$, de acordo com o procedimento normal de linearização, são calculadas a partir do modelo inicial, e ainda a partir dos parâmetros focais do sismo e do modelo exterior. O sistema (6.9) é naturalmente alargado de modo a englobar os resultados da observação de vários sismos.

A equação vectorial (6.9) presta-se à aplicação das técnicas de inversão atrás discutidas, como se evidencia escrevendo-a na forma

$$\mathbf{t}^* = G\,\mathbf{m} + \boldsymbol{\varepsilon} \qquad \text{com} \quad \mathbf{t}^* = \mathbf{t} - (\bar{\mathbf{t}} + \bar{\boldsymbol{\tau}}) \tag{6.10}$$

Dois erros que podem afectar esta equação são i) erros no modelo exterior, que afectam a parcela $\bar{\mathbf{t}}$ e ii) erros no modelo inicial, que afectam os elementos da matriz G. Os primeiros podem ser suavizados se os epicentros utilizados tiverem uma boa cobertura azimutal, bem como em distância, em relação à zona em estudo. De qualquer modo, a sua contribuição para (6.10) é da mesma ordem que a dos erros na localização epicentral, e deve ser encarada com cuidado. Os segundos erros apresentam uma contribuição de ordem superior, e podem ser absorvidos por $\boldsymbol{\varepsilon}$.

Esta metodologia foi aplicada a um conjunto de dados sísmicos registados na região da Estremadura entre Julho de 1987 e Junho de 1988. Os dados foram obtidos através da operação da Rede de Estações Sísmicas Temporárias da Estremadura (RESTE), um projecto que envolveu a participação do INMG e do Departamento de Ciências Geológicas da Universidade de Durham (Reino

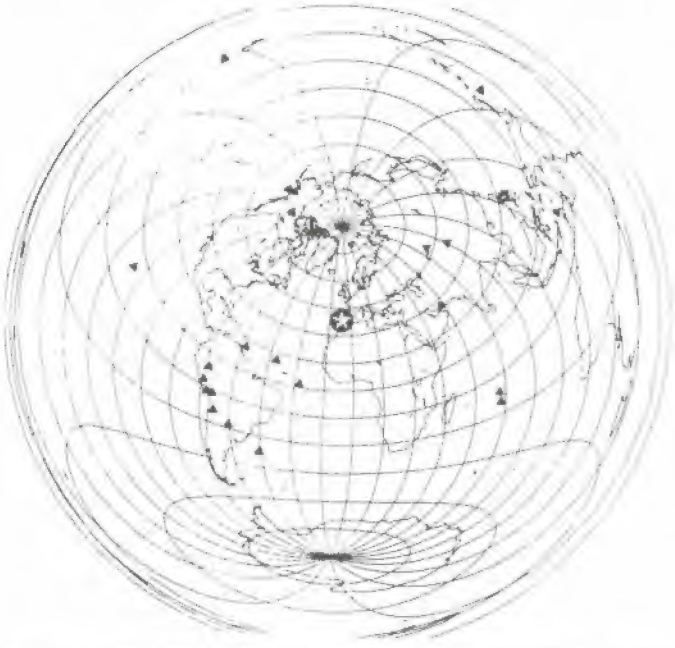


Figura 2 - Distribuição dos epicentros dos tele-sismos usados para o estudo tomográfico da litosfera da Estremadura. Os epicentros estão assinalados por triângulos; a localização da rede de estações está indicada por uma estrela. Projecção equidistante.

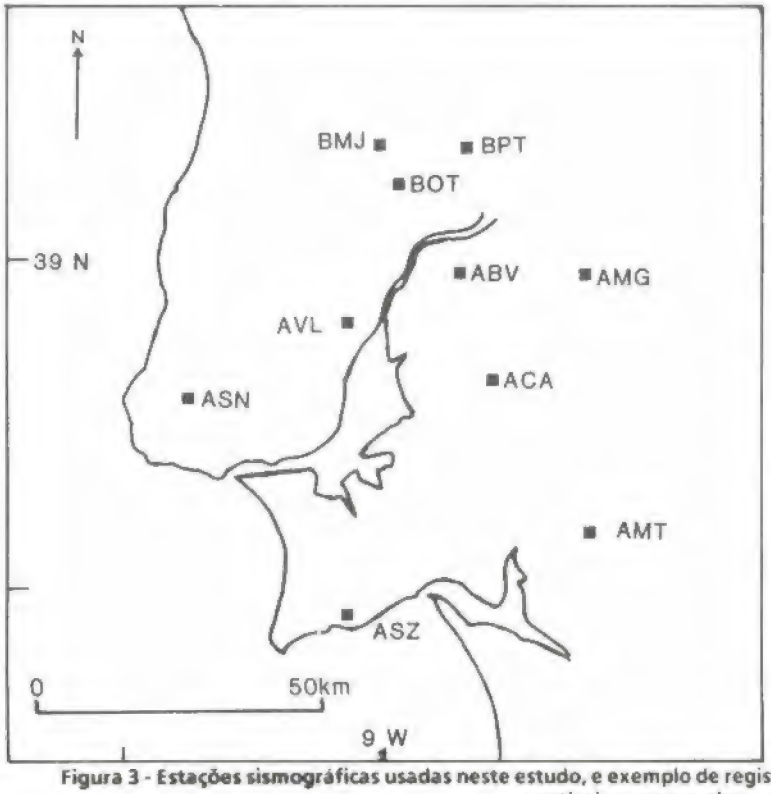
Unido). O projecto tinha por finalidade o registo de micro-sismos locais, mas foi igualmente registado um conjunto de sismos distantes, adequado ao estudo tomográfico da litosfera sob a rede de observação.

A figura 2 mostra a distribuição geográfica dos epicentros dos 20 sismos utilizados, e a figura 3 mostra o conjunto de estações sismográficas envolvidas e um exemplo de registo sísmico digital. Os dados consistiram em 140 tempos de chegada, lidos com um erro inferior a ± 0.04s.

O sistema de equações (6.10) foi construido segundo os passos já descritos, e adoptou-se para a inversão o algoritmo de inversão estocástica (4.13). Após vários ensaios, em que a resolução e a variância da solução obtida eram inspecionadas, adoptou-se como modelo de blocos o representado na figura 4. Os números aí mostrados indicam quantos raios atravessam cada bloco, e um circulo em torno do número indica que o bloco correspondente foi efectivamente modelado. O número de blocos modelados foi 59, tendo-se tomado como mínimo aceitável para a resolução de cada bloco o valor 0.70 (para o elemento correspondente da diagonal da matriz de resolução). Para o factor de amortecimento do método estocástico (coeficiente ε na equação 4.13) adoptou-se o valor 0.10, após análise do gráfico apresentado na figura 4 (*trade-off* de resolução e erro).

A figura 5 ilustra a solução obtida para as anomalias de velocidade na litosfera da Estremadura. A figura 6A mostra com mais detalhe a solução correspondente à camada mais superficial (0 a 30 km), aproximadamente coincidente com a crusta. Os valores numéricos correspondem ao desvio percentual da velocidade dos blocos em relação à média da camada correspondente. A distribuição dos pontos reflecte o facto de que as fronteiras entre os blocos foram orientadas em três direcções diferentes, sendo as várias soluções fundidas. Os desvios-padrão das anomalias obtidas não ultrapassam 0.75%, pelo que são tomadas como significativas as flutuações superiores a 2.5%. Na figura, as anomalias positivas são assinaladas por H, e as anomalias negativas assinaladas por L. A figura 6B apresenta o resultado da integração vertical das anomalias para toda a espessura da litosfera (0 a 150km). A discussão do significado geológico das anomalias observadas pode ser encontrado em [8].

NOTA: Este artigo tem por base uma monografia realizada pelo autor em 1983 no âmbito da cadeira de



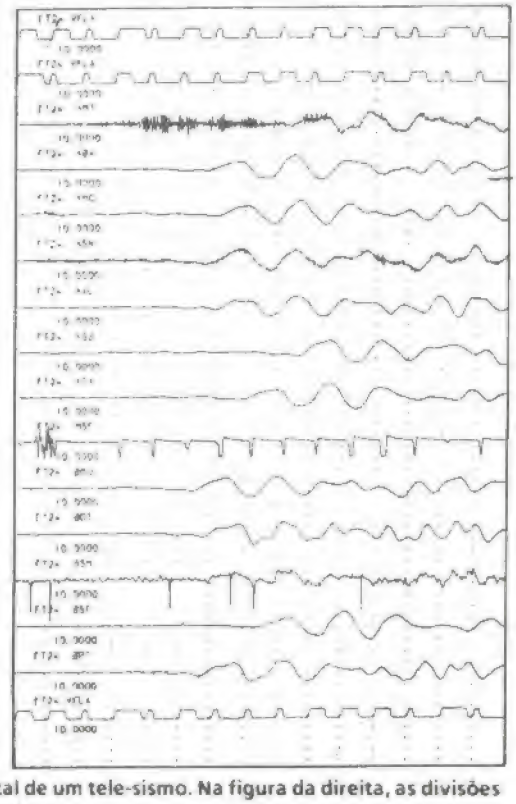


Figura 3 - Estações sismográficas usadas neste estudo, e exemplo de registo digital de um tele-sismo. Na figura da direita, as divisões verticais correspondem a 1 segundo.

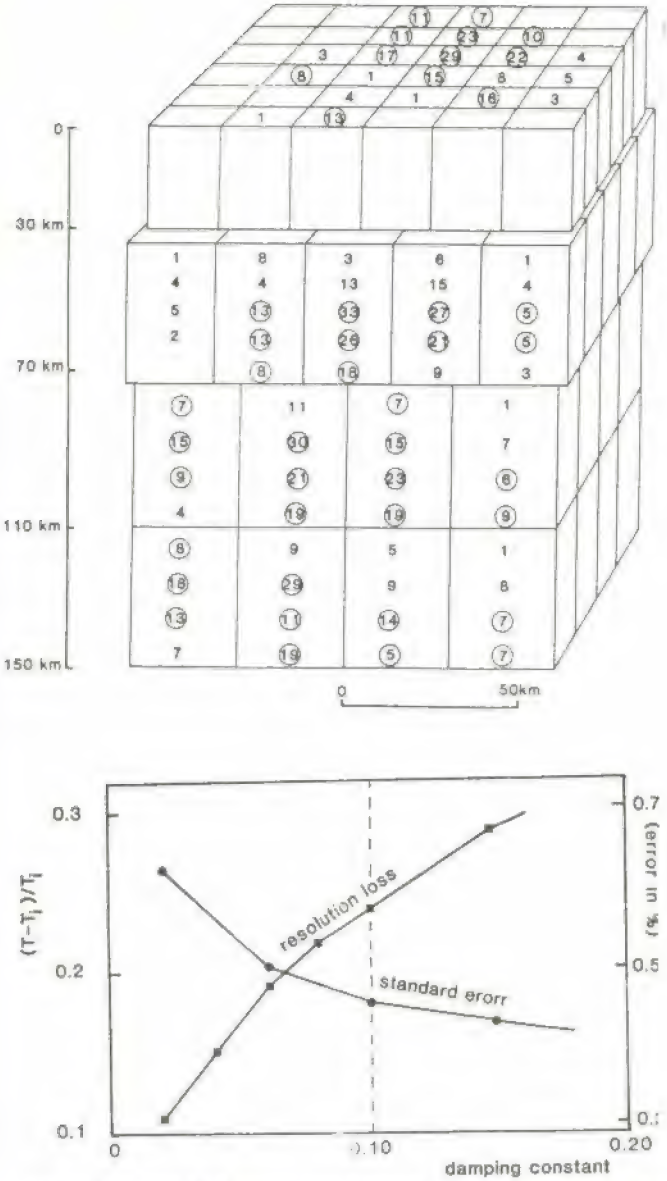


Figura 4 - Modelo de blocos usado para a litosfera da Estremadura (assume-se que a velocidade das ondas sismicas é constante no interior de cada bloco). Em baixo, mostra-se a variação do desvio padrão médio e da perda de resolução em função do parâmetro ε do algoritmo de inversão estocástica. Este diagrama é um exemplo de trade-off entre resolução e erro. T é o traço da matriz de resolução, T_I é o traço da matriz identidade da mesma ordem.

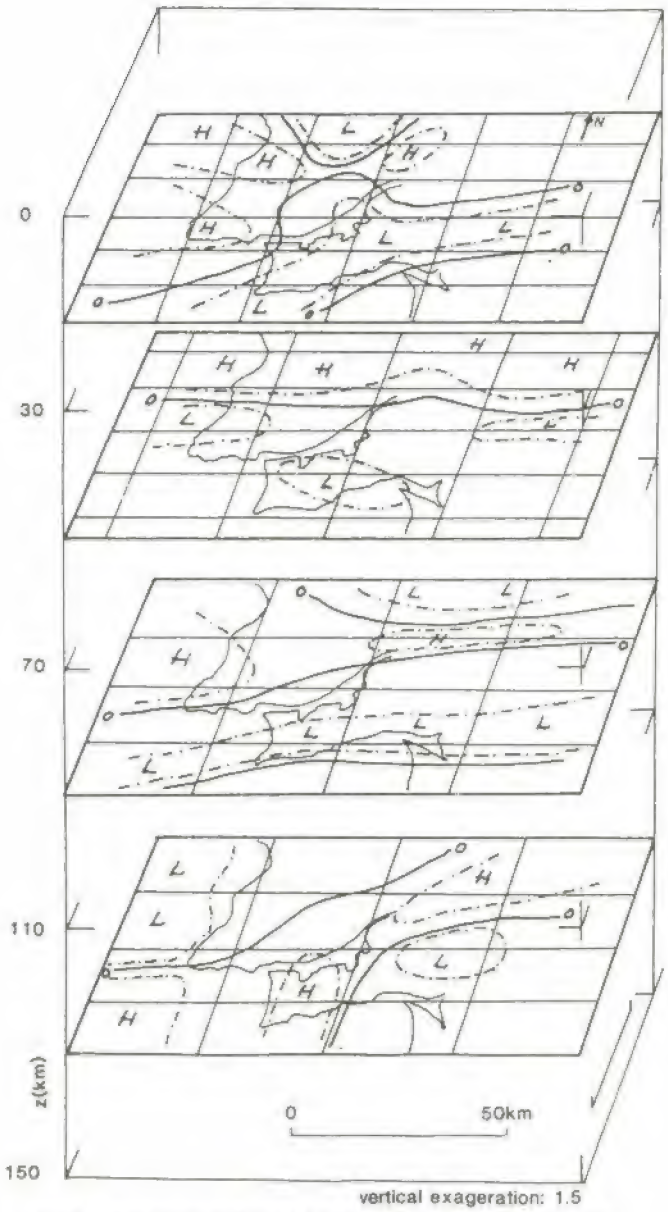


Figura 5 - Representação gráfica da solução adoptada para as anomalias de velocidade sismica na litosfera da Estremadura. Indicam-se por L as regiões de anomalia negativa em relação à media da camada, e por H as regiões de anomalia positiva. A separação entre isolinhas é de 2.5%. A discussão do possivel significado geologico das anomalias pode ser encontrada em[8].

Prospecção Geofísica da Faculdade de Ciências de Lisboa. A aplicação descrita na Secção 6 foi levada a cabo pelo autor no Departamento de Ciências Geológicas da Universidade de Durham (Reino Unido), como bolseiro da Fundação Gulbenkian.

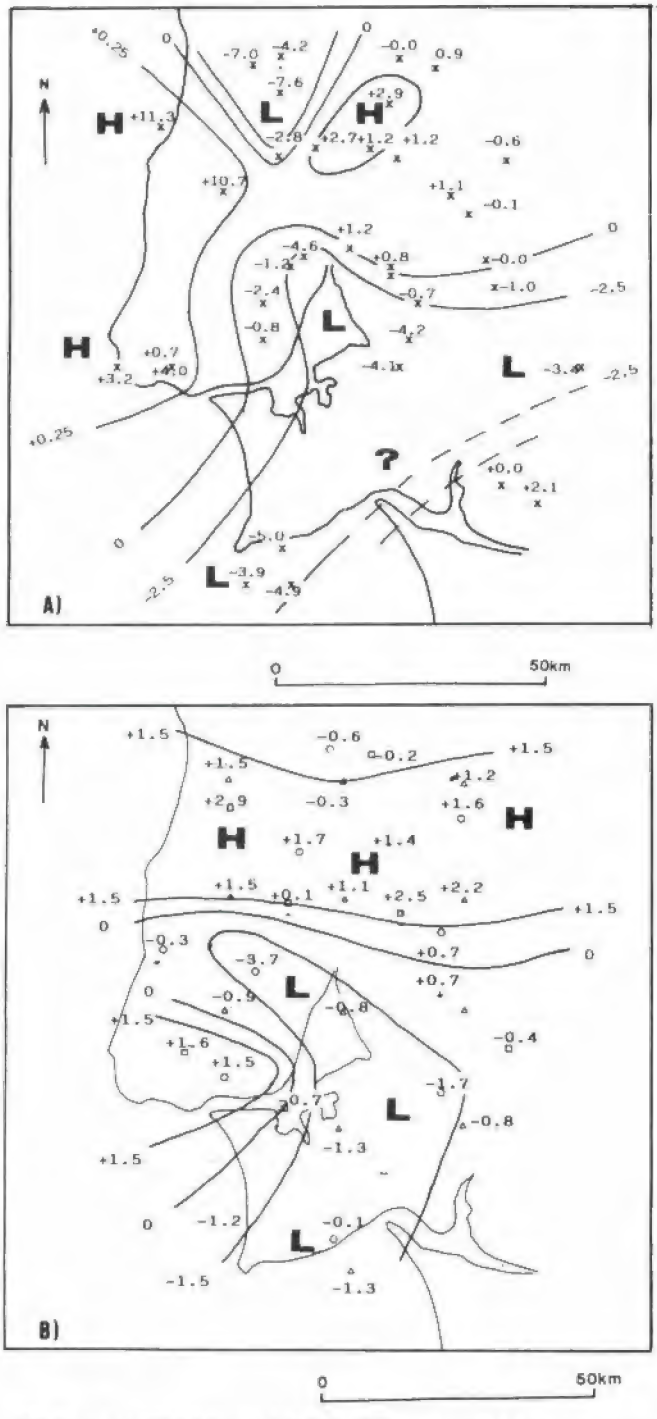


Figura 6 - A) Pormenor da solução adoptada para a primeira camada do modelo, correspondente à crosta. Os valores numéricos representam desvios em relação à velocidade sísmica média da camada, em percentagem. B) Média vertical das anomalias de velocidade, em percentagem, de 0 a 150 km de profundidade.

REFERÊNCIAS

[1] Backus, G., Gilbert, J. (1967). Numerical Applications of a Formalism for Geophysical Inverse Problems, *Geoph. Journ. R. Astr. Soc.*, **13**.

[2] Backus, G., Gilbert, J. (1968). The Resolving Power of Gross Earth Data, *Geoph. Journ. R. Astr. Soc.*, **16**.

[3] Tarantola, A. (1987). *Inverse Problem Theory: Methods for data fitting and model parameter estimation*, Elsevier, The Hague.

[4] Aki, K., Richards, P. (1980). *Quantitative Seismology - Theory and Methods (Vol. I e II)*, Ed. W.M. Freeman and Company, S.Francisco.

[5] Lanczos, C. (1961). *Linear differential operators*, D. Van Nostrand Company Ltd., London.

[6] Parker, R. (1977). Understanding Inverse Theory, *Ann. Rev. Earth Planet. Sci.*, **5**.

[7] Aki, K., Christofferson, A., Husebye, E. (1977). Determination of the Three Dimensional Seismic Structure of the Lithosphere, *J. Geoph.Res.*, **82**, 2.

[8] Fonseca, J.F.B.D. (1989). *Seismicity and Regional Tectonics of the Estremadura, SW Portugal*, tese de doutoramento, Universidade de Durham, Ed. Dep. Física do I.S.T.

BIBLIOGRAFIA ADICIONAL

[1] Bleistein, N., Cohen, J. (1982). The Velocity Inversion Problem - Present Status, New Directions, *Geophysics*, **47**, 11.

[2] Clearbout, J. (1976). *Foundamentals of Geophysical Data Processing*, Ed. McGraw Hill, N. York.

[3] Cuer, M., Grasson, J., Pascal, G. (-). Comparison Between Two Inverse Techniques (não publicado).

[4] Jackson, D. (1979). The use of a priori Data to Resolve Non-Uniqueness in Linear Inverse Problems, *Geoph. Journ. R. Astr. Soc.*, **57**.

[5] Raz, S. (1982). A Procedure for Multidimensional Inversion of Seismic Data, *Geophysics*, **47**, 10.

[6] Robinson, E., Treitel, S. (1980). *Geophysical Signal Analysis*, Ed. Prentice Hall, Englewood Cliffs.

[7] Sabatier, P. (1977). Positivity Constraints in Linear Inverse Problems I e II, *Geoph. Journ. R. Astr. Soc.*, **48**.

[8] Tarantola, A., Valette, B. (1980). The Parameterized Non-Linear Inverse Problem in Geophysics (não publicado).

[9] Treitel, S., Lines, L. (1982). Liner Inverse Theory and Deconvolution, *Geophysics*, **47**, 8.

[10] Ursin, B. (1981). Error Estimates for Inverse Modeling Schemes Using Seismic Traveltimes, *Geophysics*, **47**, 8.

[11] Weglein, A., Boyse, W., Andersen, J. (1981). Obtaining Three-Dimensional Velocity Information Directly from Reflection Seismic Data: an Inverse Scattering Formalism, *Geophysics*, **46**, 8.

The ISFET in Medicine

C. A. C. Sequeira

Laboratório de Electroquímica, Instituto Superior Técnico

SUMÁRIO

Referem-se, muito resumidamente, as aplicações dos ISFETs em medicina.

ABSTRACT

This short paper summarizes the great significance of ISFETs in medicine.

The emergence of a new type of solid electrodes, utilizing the electrochemical properties of ion-selective membranes and the field effect of semiconductor devices based on metal-insulator-semiconductor (MIS) transistors [1,2], constitutes an important advance in all branches of science and technology where the problem of expeditious and selective determination of small amounts of substances in presence of others arises. An electrode of this type is designated ISFET (ion-selective field effect transistor) if it is used for determination of ions in solution. The first ISFET was described by Bergveld [3], in 1970, and since then many papers have appeared on this matter.

Electrolyte concentration and vital gas partial pressures in blood are widely utilized, both as indicators of well being and as significant diagnostic information. Instantaneous measurement of these parameters would be of great benefit not only to an anesthesiologist during surgery but also to a critical care nurse in providing optimum patient care.

Specific ions such as Na^+, K^+, Cl^- and Ca^{2+} are often of interest, as is pH, pO_2, and pCO_2 in monitoring the basic metabolic status of a patient. The characteristics of an ISFET make it a very

Table I - Ion-Selective Field Effect Transistors in Medicine

Ions determined	Membrane	Insulator	Comments, references
Na^+	SiO_2	SiO_2	Applications for electrophysiological purposes have been given [3,11].
	Alumino-silicate	Si_3N_4 on SiO_2 surface	Production of a micromultisensor having two functions (Na^+ and pH) simultaneously [7, 12, 13]. Used for determining the tissue pH value of fish kidney.
H^+	Si_3N_4	Si_3N_4 on SiO_2 surface	Si_3N_4 is most convenient for H^+ determination than SiO_2 [7,8,14-17]. Flow injection analysis and ISFETs are combined [15]. Sensors in catheter form have proved effective in serum and blood [16,17].
K^+	Valinomycin in PVC with plasticizer	Si_3N_4 and SiO_2	A spear-shaped sensor is described for intracellular and extracellular measurements [12]. K^+ is measured in horse serum [18].
Cl^-	$AgCl$- Ag_2S	Si_3N_4	Cl^- is measured in sweat [9]. Micromultisensors having other functions (e. g., Ag^+) have been described [17].
Ca^{2+}	Ion-exchange Ca^{2+} salt in PVC	Standard MIS transistor	A stable and sensitive construction for biomedical purposes is described [19].
	Si_3N_4	Si_3N_4 on SiO_2 surface	Micromultisensors having several functions (Ca^{2+}, pH, K^+) have been developed [4-6, 15].

attractive device for continuous in-vivo monitoring of these blood species, and indeed it has proved to be successful for such monitoring in animal studies [4-9].

The operation of ISFET sensors is similar to that of MIS transistors [10]. The only difference is that the function of the gate is performed by a sensing layer analogous to the the sensing element of ISEs. The sensing layer may be the oxide coating of the gate itself, but it is usually a coating made of a material constituing the sensing element of ISEs integrated with an insulating layer in contact with the gate oxide. Like the conventional potentiometric sensors, the ISFET is also placed into the solution to be tested together with a reference electrode. The potential arising between the sensing layer and the solution is in series with the gate of the device and as it tracks the sensing species it changes the threshold of the device. The drain current of the device is then a measurement of the ion-activity (or concentration) in solution.

The technology used to produce the basic ISFET device is similar to that utilized in commercial VLSI circuits [10], so that new or exotic fabrication techniques are not necessary in order to realize the structure. ISFET production is currently under way in a number of institutions [20, 21, 23].

Most of the ISFET devices known at present with applicability in medicine are listed in Table I. The main merits of these devices are their small size, the use of all-solid materials, the possibility of construction of multisensors, the very short response time to changes of concentration, the low output resistance and the absence in most constructions of contact conductors between the membrane and the transistor.

Despite their advantages, the ISFETs have drawbacks (the reference electrode, short lifetime, high temperature-dependence, encapsulation reliability, stability and linearity) which are of concern on optimizing them for the long term implantable applications. To minimize these difficulties, new dielectric multi-layers of SiO_2, Si_3N_4, SiC, Al_2O_3, Al_2S_3, ZrO_2 and insulating polymers, and the proposal of new ISFET configurations are being introduced [20, 22, 24, 25].

The author of the present communication worked recently with Prof. Janata at the University of Utah, where an excellent research group fabricates and characterizes new solid state microelectrochemical sensors, using up-to-date technologies and advanced physical methods. Preliminary work on the preparation and testing of the new gate dielectric materials (e.g., SiC and Diamond-like carbon), as well as investigation on a structure including an on chip reference electrode, showed that several current problems of the ISFETs can be eliminated, namely the cross-talk observed in multisensor arrays.

The foregoing summarized communication shows that ISFETs are of great significance in medicine, so that it will be worthwhile to intensify the research and development in this field.

References

[1] Janata, J., Huber, R.J., (1979), *Ion-Selective Electrode Rev.*, **1**, 31.

[2] Janata, J., Huber, R.J., (1980), *Ion-Selective Electrodes in Analytical Chemistry*, Freiser, H., ed., Plenum Press, New York, **2**.

[3] Bergveld, P., (1970), *IEEE Trans. Biomed. Eng.*, **17**, 70.

[4] McKinley, B.A., Saffle, J., Jordan, W.S., Janata, J., Moss, S.D., Westenshow, D.R., (1980), *Med. Instrum.*, **14**, 93.

[5] McKinley, B.A., Wong, K.C., Janata, J., Jordan, S.W., Westenshow, D.R., (1981), *Crit. Care Med.*, **9**, 333.

[6] McKinley, B.A., Houtchens, B.A., Janata, J., (1981), *Crit. Care Med.*, **9**, 845.

[7] Esashi, M., Matsou, T., (1978), *IEEE Trans. Biomed. Eng.*, **25**, 184.

[8] Janata, J., Moss, St.D., (1976), *Biomed. Eng.*, **11**, 241.

[9] Janata, J., Moss, St.D., (1979), *Anal. Chim. Acta*, **108**, 161.

[10] Janata, J., (1985), *Solid State Chemical Sensors*, Janata, J., Huber, R.J., eds., Academic Press, Inc., New York, pg.65.

[11] Bergveld, P., de Rooij, N.F., (1978), *Theory, Design, and Biomedical Application of Solid State Chemical Sensors*, Cheung, P.W., Fleming, D.G., Ko, W.H., Neuman, M.R., eds., CRC, Cleveland.

[12] Esashi, M., Matsuo, T., (1975), *J. Japan Soc. Appl. Phys.*, **44**, 339.

[13] Matsuo, T., Esashi, M., (1978), *Extended Abstracts of Electrochemical Society Spring Meeting*, Abstr. **83**, pg. 202, Seattle.

[14] Bos, M., Bergveld, P., Van Veen-Blaauw, A.M.W., (1979), *Anal. Chim. Acta*, **109**, 145.

[15] Ramsing, A.V., Janata, J., Ruzicka, J., Levy, M., (1980), *Anal. Chim. Acta*, **118**, 45.

[16] Comte, P., Janata, J., (1977), *Proceedings of the 30th Annual Conference on Engineering in Medicine and Biology*, Abstr. 41.3, **19**, pg. 297, Los Angeles.

[17] Vlasov, Yu. G., (1979), *J. Appl. Chem.(USSR)*, **52**, 1.

[18] Moss, S.D., Janata, J., Johnson, C.C., (1975), *Anal. Chem.*, **47**, 2238.

[19] Yee, S.S., Afromowitz, M.A., Galen, P., Anderson, T., (1978),*Extended Abstracts of Electrochemical Society Spring Meeting*, Abstr. **109**, pg. 257, Seattle.

[20] Harame, D.L., Shott, J.D., Bousse, L., Meindl, J.D., (1984), *Proceedings of the 6th Annual Conference IEEE*, pg. 444.

[21] Bousse, L., de Rooij, N.F., Bergveld, P., (1983), *IEEE Trans. on Electron Devices*, **ED-30**, 1263.

[22] Van der Shoot, B.H., Bergveld, P., Bos, M., Bousse, L.J., (1983), *Sensors and Actuators*, 4, 267.
[23] Nakajima, H., Esashi, M., Matsuo, T., (1982), *J. Electrochem. Soc.*, **129**, 141.
[24] Bousse, L., Bergveld, P., (1984), *Sensors and Actuators*, **6**, 65.
[25] Caras, S.D., Janata, J., (1985), *Anal. Chem.*, **57**, 1924.

The First Principle of Thermodynamics and the Non-Separability of the Quantities "Work" and "Heat"

Rodrigo de Abreu

Centro de Electrodinâmica, Instituto Superior Técnico
Lisboa, Portugal

RESUMO

Analisa-se um modelo constituido por dois subsistemas que interactuam através de um êmbolo, verificando-se que é impossível separar a transmissão de energia em trabalho e calor. Embora o processo seja "quasi-estático" não é reversível. O Primeiro Princípio $dU = dW + dQ$, introduzido por Clausius, não permite com generalidade uma identificação com significado físico de dW e dQ embora $dU = -pdV + TdS$ se verifique numa trajectória constituida por pontos de equilíbrio de cada subsistema.

ABSTRACT

The impossibility of separating into work and heat the energy transmitted between two subsystems through a movable piston is analysed in this article. The process here described, although "quasi-static", is not reversible. It is shown that the First Principle, $dU = dW + dQ$, introduced by Clausius, does not generally allow a physical identification of dW and dQ, although $dU = -pdV + TdS$ is verified along the equilibrium points of each subsystem.

1 - Introduction

The concepts of work and heat appear associated in the First Principle of Thermodynamics, $dU = dW + dQ$. Although concepts such as work and heat historically originated prior to Clausius's work [1,2], the formalism emerging from the First Principle culminates with the introduction of the entropy through the expression $dS = dQ/T$. The success of Clausius's analysis explains the subsequent reinterpretation of the quantities work and heat involved [3] in the First Principle. This reinterpretation imposed itself by its formal success, the acceptance of which is almost universal today. The contradictions [4,5] of these reinterpretations have been analysed elsewhere [6]. In this paper one simple but fundamental aspect is analysed.

2 - The First Principle of Thermodynamics, the energy conservation principle and the concepts of work and heat

Consider the situation represented in Fig.1

Assume that on the sides 1 and 2 of the piston there are atoms of the same monoatomic ideal gas. The piston is made of a material for which the thermal conductivity is zero and the piston is initially blocked so that it cannot move. Thus, initially no energy can pass from side 1 to side 2. We take the initial temperatures, T_1 and

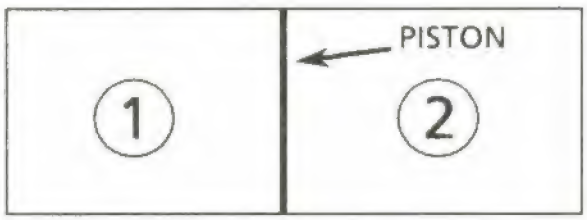


fig. 1 - An adiabatic piston surrounded by a gas. The inicial values are $P_1 = P_2$ and $T_1 \neq T_2$.

T_2, to be different, and the initial pressures, P_1 and P_2, to be equal. Once unblocked, the piston gains a translational energy to the right of order $1/2KT_1$ from a collision with a side 1 molecule, and a translational energy to the left of order $1/2KT_2$ from a collision with a side 2 molecule [7,8]. In this way energy passes mainly from side 2 to side 1 if $T_2 > T_1$.

Generally if the thermal conductivity is not zero energy can pass through the piston. But in the ideal case we are considering, the energy can only pass if the piston is moving. Since in the case the piston is "adiabatic" ("dQ" = 0 for the piston) and the process is "quasi-static" (there is equilibrium at all times on sides 1 and 2 and the pressure and temperature are well-defined on each side), one might anticipate that the process is also reversible and thus isentropic (adiabatic and reversible). In fact this is not true, as can be confirmed by solving the

Recebido para publicação em 16 de Maio de 1989.

isentropic equations,

$$P_1 V_1^{\gamma} = P_1' V_1'^{\gamma} \tag{1}$$

and

$$P_2 V_2^{\gamma} = P_2' V_2'^{\gamma} . \tag{2}$$

Here unprimed quantities refer to the initial state, primed quantities to the final state, and $\gamma = C_P/C_V$, where C_P and C_V are the specific heats at constant pressure and volume, respectively. The solution of the isentropic equations is given in the appendix.

The correct solution for the process as it actually occurs is obtained by imposing energy conservation

$$U = U_1 + U_2 = U_1' + U_2' , \tag{3}$$

equality of the inicial and final pressures on the two sides

$$P_1 = P_2 \quad , \quad P_1' = P_2' , \tag{4}$$

equality of the final temperatures on the two sides

$$T_1' = T_2' = T' \tag{5}$$

and

$$V = V_1 + V_2 = V_1' + V_2' . \tag{6}$$

The final pressure on the two sides are given by

$$P_1' = \frac{n_1}{V_1'} R T' \tag{7}$$

and

$$P_2' = \frac{n_2}{V_2'} R T' \tag{8}$$

where n_1 and n_2 are the numbers of moles of the gas in the volumes V_1' and V_2' respectively.

From Eqs. (4), (5), (6), (7), and (8) it follows that

$$V_2' = \frac{n_2}{n_1 + n_2} V \tag{9}$$

and

$$V_1' = \frac{n_1}{n_1 + n_2} V \tag{10}$$

Since the actual physical process is "quasi-static" (although not reversible), it follows that

$$dU_1 = -P_1 dV_1 + T_1 dS_1 \tag{11}$$

and

$$dU_2 = -P_2 dV_2 + T_2 dS_2 . \tag{12}$$

But, Eq.(3) implies that

$$dU_1 + dU_2 = dU = 0 . \tag{13}$$

Therefore, from Eqs. (11)-(13) it follows that

$$-P_1 dV_1 + T_1 dS_1 = P_2 dV_2 - T_2 dS_2 . \tag{14}$$

Since the pressures on the the two sides of the piston are equal at all times and the total volume is fixed

$$P_1 = P_2 \tag{15}$$

and

$$dV_1 = -dV_2 . \tag{16}$$

It follows that for the actual physical process

$$T_1 dS_1 = -T_2 dS_2 . \tag{17}$$

Of course, for an actual physical process,

$$dS = dS_1 + dS_2 > 0 . \tag{18}$$

Equations (17) and (18) immediately lead to

$$dS_1 \left(1 - \frac{T_1}{T_2} \right) > 0 . \tag{19}$$

If $T_2 > T_1$ it follows that

$$dS_1 > 0 . \tag{20}$$

From Eqs. (17) and (20) it follows that

$$dS_2 < 0 . \tag{21}$$

Moreover, if $T_2 > T_1$, then $dV_2 < 0$, $dU_2 < 0$, and $T_2 dS_2 < 0$.

In this process just considered, the pressures on the two sides of the piston are equal at all times, which means no "work" is done. However, the energy transfer occurs through the agency of the moving piston, and if one considers "work" to be energy transfered via macroscopic, non-random motion, then it appears that work is done. There is no paradox or problem here, since it is the **net** "work", the "work" done on side 1 by side 2 + the "work" done on side 2 by side 1, that is zero. Furthermore, since the thermal conductivity of the piston is zero, no "heat" flow is possible. However, the flow of energy is from the high temperature side to the low temperature side, and if one considers heat to be energy transfer resulting from a temperature difference, then it appears that there is heat flow. Obviously, there is a problem as far as what we mean by heat and work. We have here clearly, a difficulty of interpretation because we use the designations heat and work without a well defined meaning. Today the dominant ortodox point of view is that heat is what complements work in the expression $\Delta U = W + Q$, assuming that work is a well known entity with generality. This is not so because work is associated with the concept of force which has conceptual difficulties for variable mass systems [5,6].

Everything including the problem presented here has a very simple interpretation if we abandon Clausius's point of view that heat is a special kind of energy transfer (caloric conception) and only use the more abstract concept of energy. In fact if we consider **heat as internal energy** we can say that heat is transformed into work when a gas expands and the mass associated with the piston has an increased gravitational potencial energy. Of course it is possible to affirm that the gas is performing work or that heat is "transformed" into work! Consider now the well known Joule's apparatus to demonstrate the equivalence of work and heat. When the masses change their position, the gravitational potencial energy (work) is transformed into internal energy (heat). Let us consider two "heat reservoirs" at temperatures T_1 and T_2 ($T_1 > T_2$). We can link the two reservoirs by radiation and since $T_1 > T_2$, have an oriented flux of energy that we can call work because we have an oriented energy that can rise a weight! In fact, this being so, and if we define heat as energy transfer resulting from a temperature difference, then we can say that heat is work if we define work as an oriented flux of energy! Consider now one more example. A relativistic rocket moves by ejection of a photon gas. Try to define what is work and what is heat with the First Principle, $dU = dW + dQ$! The classification of energy in work and heat has only meaning in the non symmetry of the transformation of the gravitational potential energy (work) into internal energy (heat). In fact if we have the same deformation variables of a system in equilibrium we verify that work can be transformed into heat not the inverse. This being so the internal energy is not only a function of the deformation variables and we must introduce another variable which contains this non-symmetry. This variable, S, is the entropy and the non symmetry may be formally imposed by choosing S as a non decreasing function of energy. When the deformation variables are the same we have $\Delta S \geq 0$ ($(\partial U/\partial S)_V > 0$).

It should be noted that the final solution is the same whatever the pressures are in the initial state, the system going through an intermediate stage where pressures become equal evolving there upon to equal pressures and temperatures. In fact the final solution is only function of the total energy. If the thermal conductivity of the piston takes "small" values other than zero and if the piston mass is "high" enough for the transformation can be considered "quasi-static", equations (11) and (12) are still verified and final equilibrium state is the same, there existing, nevertheless, a flow of energy inside the piston. We thus conclude that the quantities values $\int_{if} P_{1,2}\, dV_{1,2}$ and $\int_{if} -T_{1,2}\, dS_{1,2}$ exist independently the thermal conductivity be or not be zero, it not being possible to associate them with quantities "work" and "heat", since the energy exchange is due to the movement of the piston and is not separated into two parts with energetic significance: work ($\int$-pdV) and heat ($\int$ TdS).

3 - Conclusion

The traditional concepts of work and heat stablished by the First Principle of Thermodynamics, $dU = dW + dQ$, are questioned through the analysis of a system constituted of a gas occupying two volumes separated by a movable dividing wall with zero thermal conductivity.

Although $dU = -pdV + TdS$ is a verified along the equilibrium points of a quasi-static transformation, and although it is possible to formally identify [9] $dW = -pdV$ and $dQ = TdS$, these quantities do not possess the physical significance commonly associated with them, in that here they do not correspond to energy exchanges with a well-defined significance [9].

As shown elsewhere [6, 10-13] and as suggested here, it is both possible and necessary to formulate thermodynamics without resorting to the First Principle.

It is therefore necessary to use the First Principle $dU = dW + dQ$ carefully, since -pdV does not always signify infinitesimal work [14-16], and TdS does not always signify heat [9].

The conceptual difficulties of thermodynamics have lately been object of special attention in the literature [14-38], and in this paper an emphasis is laid on one particular aspect which regard as fundamental [6, 10-13].

The concepts of work and heat must be related through the "Second Principle of Thermodynamics" when stating the impossibility of the internal energy (heat) [6] being converted into work in a transformation where the deformation variables are the same at the beginning and at the end of the transformation [17]. This non-simmetry implies the existence of the quantity entropy which never decreases ($\partial U/\partial S > 0$, because we choose that, or the entropy never increases and then we have $\partial U/\partial S < 0$) [10-13]. It is therefore essential to use the designations heat and work in the precise sence suggested by the "Second Principle" and not in the sence resulting from $dU = dW + dQ$ [6,37,38].

Appendix

I'm going to show that the assumption that the process considered is isentropic is an absurd.

Let us assume that the transformation is isentropic. We have

$$P_1 V_1^{\gamma} = P_1' V_1'^{\gamma}, \qquad \text{(I)}$$

$$P_2 V_2^{\gamma} = P_2' V_2'^{\gamma}. \qquad \text{(II)}$$

From (I) and (II), $P_1 = P_2$ and $P'_1 = P'_2$, we have

$$\left(\frac{V_1}{V_2}\right)^{\gamma} = \left(\frac{V_1'}{V_2'}\right)^{\gamma}. \qquad \text{(III)}$$

Since $P_1 = P_2$, $P'_1 = P'_2$, $P_{1,2}V_{1,2} = n_{1,2}RT_{1,2}$ and $P'_{1,2}V'_{1,2} = n_{1,2}RT'_{1,2}$ we have

$$\frac{V_1}{V_2} = \frac{n_1 T_1}{n_2 T_2} \text{ and } \frac{V'_1}{V'_2} = \frac{n_1 T'_1}{n_2 T'_2} \qquad \text{(IV)}$$

If $T'_1 = T'_2$ we have from (III) and (IV) the absurd that $T_1 = T_2$.

References

[1] R. B. Lindsay - Editor (1975). *Energy: Historical Development of the concept*, Benchmark Papers on Energy, Dowden, Hutchinson. & Ross, Stroudsburg, **I**.

[2] Joseph Kestin - Editor (1976). *The second law of thermodynamics*, Benchmark Papers on Energy, Dowden, Hutchinson & Ross, Stroudsburg, **V**.

[3] W.P.Allis and M.A. Herlin (1952). *Thermodynamics and Statistical Mechanics*, McGraw-Hill, New York, 67.

[4] J.E.Ahern (1980). *The exergy method of energy systems analysis*, Wiley-Interscience, New York, 43-44.

[5] C.K.Yuen (1970). *Am. J. Phys.*, **38**, 246.

[6] R. de Abreu (1983). *Contribution to the elucidation of some fundamental problems associated with the first and second principles of thermodynamics - relativistic transformations*, Ph.D. Thesis, Instituto Superior Técnico, Lisbon.

[7] R.P. Feynman, R.B. Leighton and M. Sands (1976). *The Feynman Lectures on Physics*, Addison-Wesley, Reading, Section 39, **I**, 6.

[8] A. Einstein (1909). *Physikal. Zeitschrift*, 10, 817.

[9] H.B. Callen (1960). *Thermodynamics*, Wiley, New York, 62.

[10] R. de Abreu (1985). *Técnica*,1 and 2, 100.

[11] R. de Abreu (1986). *Técnica*,3 and 4, 55.

[12] R. de Abreu (1987), *Técnica*, 1, 50.

[13] R. de Abreu (1987). The Ideal Gas Equation p=αu and the Second Principle of Thermodynamics (submitted to publication).

[14] W. Hoover and B. Moran (1979). *Am. J. Phys.*, **47**, 10.

[15] Daniel Kivelson and Irwin Oppenheim (1966). *Journal of Chemical Education*, **43**, 5.

[16] Gabriel Laufer (1983). *Am. J. Phys.*, **51**, 1.

[17] Pau-Chang Lu (1982). *Am. J. Phys.*, **50**, 3.

[18] T.W. Marshall (1978). *Am. J. Phys.*, **46**, 2.

[19] S.G. Canagaratna (1978). *Am. J. Phys.*, **46**, 12.

[20] M. Bailyn (1985). *Am. J. Phys.*, **53**, 11.

[21] J. Copeland (1982). *Am. J. Phys.*, **50**, 7.

[22] G. Falk, F. Herrmann, and G. Bruno Schmid (1983). *Am. J. Phys.*, **51**, 12.

[23] P.T. Landsberg (1983). *Am. J. Phys.*, **51**, 9.

[24] Brune Arne Sherwood and W.H. Bernard (1984). *Am. J. Phys.*, **52**, 11.

[25] Jack Copeland (1982). *Am. J. Phys.*, **50**, 7.

[26] J. de Heer (1977). *Am. J. Phys.*, **45**, 12.

[27] Harvey S. Leff (1978). *Am. J. Phys.*, **46**, 3.

[28] H. Erlichson (1984). *Am. J. Phys.*, **52**, 7.

[29] N.G. Van Kampen (1968). *Phys Rev.*,**173**.

[30] W.H. Bernard (1984). *Am. J. Phys.*, **52**, 3.

[31] H.R. Kemp (1985). *Am. J. Phys.*, **53**, 10.

[32] Brune Arne Sherwood (1983). *Am. J. Phys.*, **51**,7.

[33] H. Callen and G. Horwitz (1971). *Am. J. Phys.*, **39**,8.

[34] P.T. Landsberg (1980). *Physical Review Letters*, **45**, 3.

[35] P.A. Goodinson and B. Luffman (1980). *Il Nuovo Cimento*, **60B**, 1.

[36] P.F. González-Diáz (1981). *Il Nuovo Cimento*, **64B**, 2.

[37] J. Serrin (1986). *New Perspectives in Thermodymamics*, Springer-Verlag, New York, 5.

[38] Barrow, G. M. (1988). *J. Chem. Educ.*, **65**, 122.

Dual Galilei Transforms

M. J. Martins

Centro de Electrodinâmica da Universidade Técnica de Lisboa (I.N.I.C.)

RESUMO

A electrodinâmica de meios em movimento relativista é analisada usando uma transformação de coordenadas dual da de Galileu.
Partindo duma transformação semelhante à de Lorentz, e tomando o limite $v/u^2 \to \infty$, obtêm-se as transformações duais de Galileu. Isto permite reduzir a equação de onda à forma hiperbólica usual.

ABSTRACT

Using a coordinate transformation dual of the Galilei transform we obtain a significant simplification in the determination of the fields in moving media.
We departed from a Lorentz like transform where by taking the limit $v/u^2 \to \infty$ we obtain a dual Galilei transform..This reduces the equations of the fields in moving media to the usual hyperbolic form.

Vacuum electrodynamics with external changes and currents may be described in a simple covariant manner.

In moving material media, constitutive relations do not, in general, assume a simple form. However for many problems of radiation one may assume that the index of refraction $n=(\mu\varepsilon/\mu_0\varepsilon_0)^{1/2}$ remains constant over some range of frequencies. This approach has been widely used to study radiation in material media [1, 8, 5, 7].

In fact it has been assumed that the equations pertaining to vacuum electrodynamics may still be used, with the velocity of light c replaced by $u=c/n$, phase velocity of waves in the medium.

Problems involving uniformly moving charges with $v<u$ may be reduced to the static case by this transformation.

$$r'_v=\gamma_1(r-vt)_v \qquad r'_t=r_t \qquad t'=\gamma_1\left(t-\frac{v.r}{u^2}\right)$$
$$J'_v=\gamma_1(J-\rho V)_v \qquad J'_t=J_t \qquad \rho'=\gamma_1\left(\rho-\frac{v.r}{u^2}\right) \tag{1}$$
$$\gamma_1=\left(1-\frac{v^2}{u^2}\right)^{-\frac{1}{2}}$$

Although this transformation keeps Maxwell's equation and $\mathbf{r}\cdot\mathbf{r}-u^2t^2$ invariant, it has no physical meaning. It is only a useful mathematical aid for problem solving purposes.

For $v>u$ this transformation becomes imaginary [4].It is then convenient to find a similar transformation which might simplify the derivation of the fields for $v>u$.

By replacing v by u^2/v in (1) one obtains;

$$r'_v=\gamma_2\left(r-\frac{u^2}{v}t\right)_v \qquad r'_t=r_t \qquad t'=\gamma_2\left(t-\frac{v.r}{v^2}\right)$$
$$J'_v=\gamma_2\left(J-\frac{u^2}{v}\rho\right)_v \qquad J'_t=J_t \qquad \rho'=\gamma_2\left(\rho-\frac{v.J}{v^2}\right) \tag{2}$$
$$\gamma_2=\left(1-\frac{u^2}{v^2}\right)^{-\frac{1}{2}}$$

By using this transformation, the charge density disappears and we are led to a simple antenna problem. As before, this transformation keeps Maxwell's equations and $\mathbf{r}\cdot\mathbf{r}-u^2t^2$ invariant. When $v/u\to 0$ transformation (1) reduces to Galilei's.

By taking the opposite limit $v/u^2\to\infty$ in (2) we obtain

$$r'_v=r_v \qquad r'_t=r_t \qquad t'=t-\frac{\mathbf{v.r}}{v^2} \tag{3}$$

which is similar to a Galilei transform by interchanging space and time coordinates. We call it a "Dual Galilei Transform".

As an example we will apply it to the determination of the fields in a moving homogeneous medium.

Departing from Maxwell's equations and Minkowsky constitutive relations for a moving medium [4].

$$\mathbf{D}=\varepsilon'a\mathbf{E}+\Omega\left(\mathbf{v}\times\mathbf{H}-\varepsilon'\mathbf{v}(\mathbf{v}.\mathbf{E})\right)$$
$$\mathbf{B}=\mu'a\mathbf{H}-\Omega\left(\mathbf{v}\times\mathbf{H}+\mu'\mathbf{v}(\mathbf{v}.\mathbf{H})\right) \tag{4}$$

where

$$\mathbf{v}=\frac{\mathbf{v}}{r} \qquad \boldsymbol{\beta}=\frac{\mathbf{v}}{c}$$

$$a=\frac{1-\beta^2}{1-n^2\beta^2} \tag{5}$$

$$\boldsymbol{\Omega}=\frac{(n^2-1)}{1-n^2\beta^2}\,\frac{\boldsymbol{\beta}}{c} \tag{6}$$

we may write

$$\begin{aligned}
&\nabla_n\times\mathbf{H}-\varepsilon' a\frac{\partial\mathbf{E}_n}{\partial t}=\mathbf{J}\\
&\nabla_n\times\mathbf{E}+\mu' a\frac{\partial\mathbf{H}_n}{\partial t}=0\\
&\nabla_n\cdot\ \mathbf{E}_n=\frac{1}{\varepsilon' a}(\rho+c\,\boldsymbol{\Omega}.\mathbf{J})\\
&\nabla_n.\mathbf{H}_n=0
\end{aligned} \tag{7}$$

where

$$\begin{aligned}
&\nabla_n=\nabla-\boldsymbol{\Omega}\frac{\partial}{\partial t}\\
&\mathbf{E}_n=\mathbf{E}-\mathbf{v}(\mathbf{v}.\mathbf{E})\frac{(a-1)}{a}\\
&\mathbf{H}_n=\mathbf{H}-\mathbf{v}(\mathbf{v}.\mathbf{H})\frac{(a-1)}{a}
\end{aligned} \tag{8}$$

Owing to the anisotropy caused by the movement, the above equations separate into transverse and parallel components.

The transverse part is

$$\begin{aligned}
&\left\{\left(\mathbf{v}.\nabla_n\right)^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial t^2}\right\}(\mathbf{v}\times\mathbf{H})=\\
&\mathbf{v}\times\left\{\nabla\left(\mathbf{v}.\nabla_n\right)(\mathbf{v},\mathbf{H})-\mathbf{v}\times\left[\varepsilon' a\,\nabla\frac{\nabla\partial}{\partial t}(\mathbf{v}.\mathbf{E})+\left(\mathbf{v}.\nabla_n\right)\mathbf{J}\right]\right\}
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\left\{\left(\mathbf{v}.\nabla_n\right)^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial t^2}\right\}(\mathbf{v}\times\mathbf{E})=\\
&\mathbf{v}\times\left\{\nabla\left(\mathbf{v}.\nabla_n\right)(\mathbf{v}.\mathbf{E})+\mu' a\frac{\partial}{\partial t}\left[\mathbf{v}\times\nabla(\mathbf{v}.\mathbf{H})-\mathbf{J}\right]\right\}
\end{aligned} \tag{10}$$

Whereas the parallel part takes the form

$$\begin{aligned}
&D_n(\mathbf{v}.\mathbf{E})=\frac{1}{\varepsilon' a}+\left\{\left(\mathbf{v}.\nabla_n\right)P_n+\frac{a^2}{u^2}\frac{\partial}{\partial t}(\mathbf{v}.\mathbf{J})\right\}\\
&D_n(\mathbf{v}.\mathbf{H})=-\mathbf{v}\times\nabla\,\mathbf{J}
\end{aligned} \tag{11}$$

and

$$D_n=\mathbf{v}\times\nabla.\mathbf{v}\times\nabla+\frac{1}{a}\left(\mathbf{v}.\nabla_n\right)^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial\gamma^2} \tag{13}$$

$$P_n=P-c\,\boldsymbol{\Omega}.\mathbf{J} \tag{14}$$

By introducing the Dual Galilei Transformation

$$\begin{aligned}
&\mathbf{r}'=\mathbf{r} \qquad t'=t-\boldsymbol{\Omega}.\mathbf{r}\\
&\frac{\partial}{\partial t}=\frac{\partial}{\partial t'}
\end{aligned} \tag{15}$$

$$\left(\mathbf{v}.\nabla-\Omega\frac{\partial}{\partial t}\right)=(\mathbf{v}.\nabla') \tag{16}$$

which gives

$$D_n=\mathbf{v}\times\nabla'.\mathbf{v}\times\nabla'+\frac{1}{a}\left\{\left(\mathbf{v}.\nabla'\right)^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial t'^2}\right\} \tag{17}$$

$$\left(\mathbf{v}.\nabla-\Omega\frac{\partial}{\partial\gamma}\right)^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial t^2}=(\mathbf{v}.\nabla')^2-\frac{a^2}{u^2}\frac{\partial^2}{\partial t^2} \tag{18}$$

So, the operators are expressed in their normal hyperbolic form.

The solutions of the equations (9), (10), (11) and (12) are easily obtained by solving for (**v**.**E**) and (**v**.**H**) in (11) and (12), in the new reference frame.

The solution to the problem is found by introducing the results in equations (9) and (10) and transforming back to the original system.

References

[1] Panofsky, W., Phillips, M. (1972). *Classical Electricity and Magn etism*, Addison Wesley.

[2] Faro, M. (1980). *Acad. Cienc. Lisboa* , **XXIII**.

[3] Fox, R., Lipman, S. (1986). *Proc. Roy. Soc.*, A 416.

[4] Kong, J.A. (1972). *Theory of Electromgn. Waves*,Wiley.

[5] Lerche, I. (1982). *Astr. Phys. Journ.* , **160**, 1003.

[6] Martins, M.J. (1987). *Técnica*, 1 , 97.

[7] Martins, M.J. (1988). *Técnica*, 2 , 62.

ADIST - Associação para o Desenvolvimento do Instituto Superior Técnico

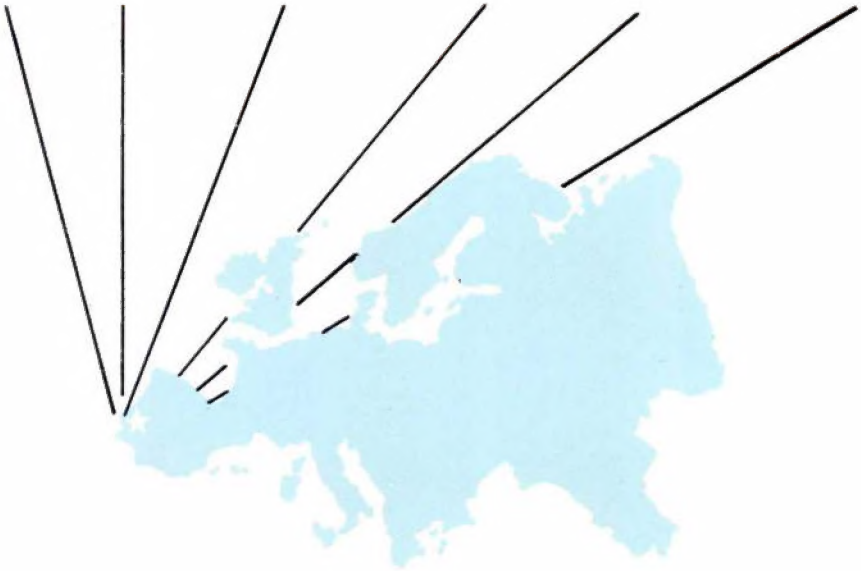
- **Formação Especializada**
- **Apoio a projectos de Investigação e Desenvolvimento**

Uma Associação reconhecida como de Utilidade Pública

ao serviço do Instituto Superior Técnico

Av. Rovisco Pais , 1096 Lisboa Codex - Telef. 802045

O IST EM PORTUGAL E NA EUROPA
LEVAMOS O ENSINO MAIS LONGE



LICENCIATURAS

Eng.ª dos Materiais
Eng.ª de Minas
Eng.ª Civil
Eng.ª Electrotécnica e de Computadores
Eng.ª Mecânica
Eng.ª Química
Física-Tecnológica
Matemática Aplicada e Computação
Eng.ª Naval
Eng.ª Informática e de Computadores
Eng.ª e Gestão Industrial

MESTRADOS

Matemática Aplicada
Eng.ª Electrotécnica e de Computadores
Transportes
Eng.ª de Estruturas
Construção
Eng.ª Mecânica
Física e Eng.ª dos Plasmas
Eng.ª Química (Processos e Indústria)
Hidráulica e Recursos Hídricos
Mineralurgia e Planeamento Mineiro
Planeamento Regional e Planeamento da Energia
Física das Altas Energias, Nuclear e Instrum.
Química dos Processos Catalíticos
Eng.ª Química (Química Aplicada)
Biotecnologia (Eng.ª Bioquímica)
Investigação Operacional e Eng.ª de Sistemas
Eng.ª de Materiais
Física
Ciência e Tecnologia de Alimentos

DOUTORAMENTOS

Eng.ª dos Materiais
Eng.ª de Minas
Eng.ª Civil
Eng.ª Electrotécnica e de Computadores
Eng.ª Mecânica
Eng.ª Química
Física-Tecnológica
Matemática Aplicada e Computação
Eng.ª Naval
Eng.ª de Sistemas
Eng.ª Física
Eng.ª e Gestão Industrial
Eng.ª Informática e de Computadores
Química